POLITECNICO DI MILANO
Corso di Laurea in Ingegneria Informatica
Dipartimento di Elettronica e Informazione

ASSEGNAMENTO AUTOMATICO DI MACROCATEGORIE AGLI ARTICOLI DI WIKIPEDIA

Relatore: Prof. Marco Colombetti
Correlatori: Ing. Riccardo Tasso, Ing. David Laniado

Tesi di Laurea di:
Jacopo Farina, matricola 713091

Anno Accademico 2009-2010

Ai miei genitori,
che hanno reso possibile questo lavoro
(e molte altre cose)
con i loro sacrifici

Sommario

Wikipedia è un' enciclopedia online, fondata 2001, il cui contenuto è liberamente modificabile da ogni utente. Ogni pagina di Wikipedia è assegnata a una o più categorie, definite anch'esse dagli utenti, che sono a loro volta assegnate a una o più categorie, dando origine a una struttura che pur essendo generalmente tassonomica può contenere cicli e anomalie di tutti i tipi.
Lo scopo della tesi è individuare dei criteri per scegliere, tra un insieme di categorie generiche scelte arbitrariamente, quella più adatta a contenere un certo articolo di Wikipedia, utilizzando il grafo degli assegnamenti alle categorie.
Sono stati analizzati nove criteri di assegnamento, di cui uno ideato e usato in passato da alcuni studiosi, e per ognuno è stata valutata la similitudine tra i risultati ottenuti automaticamente e quelli stabiliti manualmente da un valutatore umano, verificando l'efficacia dei diversi criteri e ottenendo molti dati statistici sulla struttura di Wikipedia.

Ringraziamenti

Ringrazio Riccardo Tasso e David Laniado, per la grande disponibilità mostrata in tutte le fasi del lavoro e per avermi dato molta autonomia.
Ringrazio inoltre gli utenti dell'edizione di Wikipedia in lingua inglese per la descrizione dell'algoritmo di Tarjan per le strutture fortemente connesse e della cosine similarity.
Ringrazio i miei genitori, che non mi hanno fatto mai mancare nulla standomi sempre vicini con il loro affetto, nel faticoso compito di crescermi e educarmi.
Ringrazio mia nonna Rosetta, che si è presa cura di me durante l'infanzia in tutti i modi possibili, regalandomi tanti ricordi di pranzi meravigliosi, pomeriggi ai giardini e molte altre cose che non dimenticherò mai.

Indice

Capitolo 1

Introduzione

Wikipedia è un' enciclopedia online, fondata nel gennaio 2001[11], il cui contenuto è liberamente modificabile da ogni utente. Chiunque si colleghi al sito può creare nuovi articoli o modificare quelli già esistenti.

Dalla sua fondazione il progetto ha avuto una crescita rapida e costante, e si è assistito allo sviluppo di una comunità di editori abituali che ha stabilito delle linee guida sempre più precise e si è organizzata in *progetti* che si occupano di aspetti specifici dell'enciclopedia.

Il successo di Wikipedia ha incoraggiato lo sviluppo di progetti simili volti a creare enciclopedie su argomenti specifici o parodistiche.

Per molti versi Wikipedia è assimilabile a un campione rappresentativo dell'intero web, in quanto copre praticamente ogni argomento, ha numerosissimi editori indipendenti che immettono informazioni in quantità e qualità molto diverse tra di loro senza un controllo centrale, le pagine possono essere estremamente lunghe e dettagliate o molto scarne, presentare numerosi collegamenti o non presentarne affatto e in generale non c'è uno standard qualitativo a cui si è obbligati a tenersi, nonostante esistano delle linee guida e delle indicazioni seguite dalla maggior parte degli utenti.

Come avviene sempre più frequentemente sul web, anche in Wikipedia è sorto il problema di organizzare i dati in maniera organica, ed è stato affrontato tramite l'introduzione, nel maggio del 2004[10], delle categorie.

Le categorie sono degli insiemi di articoli con delle similitudini relative ai contenuti, per esempio ci sono categorie che contengono tutti gli articoli sui comuni della Lombardia o sui premi Nobel per la chimica, e possono a loro volta essere contenute in altre categorie e contenerne altre.

La struttura delle categorie non è strettamente gerarchica, perché l'assegnamento di una categoria a un'altra è effettuato anch'esso liberamente dai singoli utenti, sulla base della similitudine degli argomenti, quindi possono

esserci categorie che si contengono a vicenda, categorie non assegnate a nessun altra o vuote, nonostante questi casi siano visti come anomalie e quindi scoraggiati dagli utenti attivi del sito.

L'enciclopedia è disponibile in numerose lingue, la versione in una lingua è indipendente dalle altre quindi possono esserci articoli solo in certe lingue o con contenuti e lunghezze diverse a seconda dei casi. La versione in lingua inglese è quella più grossa in termini di voci, lunghezza degli articoli e numero di categorie, ed è quindi quella che verrà presa in esame.

Il problema che verrà affrontato in questa tesi è quello di trovare una tecnica per assegnare automaticamente un articolo di Wikipedia alla categoria più adatta a contenerlo tra quelle in un insieme prestabilito, discutendo vari metodi per effettuare questa operazione e valutare la qualità dei risultati e alcuni utilizzi molto semplici degli stessi per ottenere delle informazioni statistiche sui contenuti del progetto.

Le categorie appartenenti all'insieme delle categorie prestabilite saranno chiamate *macrocategorie*.

Il problema è interessante perché, con l'aumento esponenziale della quantità di pagine disponibili sul web, l'esigenza di strumenti più potenti per organizzare, mettere in relazione e ricercare più efficacemente i dati diventa più forte.

Potendo svolgere automaticamente questa operazione si avrebbe uno strumento utile per molte applicazioni.

Ad esempio sarebbe possibile recuperare tutti gli articoli relativi a un certo argomento per condurre delle analisi statistiche sulla struttura di Wikipedia, in modo da capire quali argomenti sono oggetto della maggiore copertura da parte del sito e come questa copertura cambia nel corso del tempo, lavoro che risulta praticamente impossibile da svolgere a mano vista la quantità di articoli.

Inoltre, questo strumento sarebbe utile anche all'interno della comunità stessa degli editori per individuare approssimativamente l'affinità tra due utenti in termini di argomenti degli articoli modificati, allo scopo di suggerire a un utente un progetto a cui partecipare o delle pagine da inserire in un template.

Un'altra possibilità è la riorganizzazione delle categorie in una struttura diversa, per esempio una tassonomia, grazie all'applicazione ricorsiva del metodo, o dei semplici elenchi, per permettere in un secondo momento al visitatore di effettuare una ricerca di un articolo riguardante qualcosa di cui non conosce il nome ma che è in grado di assegnare alle categorie.

Molti siti, in particolare i blog, utilizzano i tag per organizzare i propri contenuti.

I tag sono delle parole chiave assegnate manualmente dall'utente che scrive un articolo al momento della sua pubblicazione, e permettono ai visitatori di passare alla lettura di altri articoli con le stesse parole chiave suggeriti automaticamente dalle piattaforme di blogging.

In altri casi, come avviene con le foto su Facebook, i tag sono assegnati da molti utenti nel corso del tempo, oltre che dall'utente che pubblica il contenuto, costituendo così un esempio di organizzazione dei contenuti effettuata in maniera distribuita detta *folksonomy*, attualmente oggetto di numerosi studi e tentativi di utilizzo efficace da parte delle aziende.

Il problema dei tag, però, è che collegano in orizzontale degli articoli ma non permettono di creare una struttura anche solo approssimativamente gerarchica come quella delle categorie di Wikipedia.

È invece possibile fare l'opposto, ossia utilizzare le categorie di Wikipedia come se fossero dei tag ignorando le appartenenze tra categorie, quindi le categorie di Wikipedia possono definire una struttura più espressiva rispetto a un sistema di tagging. Questi dati potrebbero quindi essere utilizzati da un' applicazione per la gestione dei contenuti che integri l'organizzazione a tag e quella a categorie.

Si potrebbero così suggerire agli utenti dei termini da usare per classificare i propri articoli, basandosi sulla tendenza dei tag, o meglio degli articoli con i nomi uguali ai tag, assegnati in passato dall'utente di appartenere alle stesse macrocategorie. Se ad esempio l'utente usa spesso i tag *Google* e *Yahoo* potrebbe essere interessato all'uso del tag *Bing* o *Lycos* che sono contenuti in *Internet search engines* insieme agli altri due. Se però l'utente usa spesso tag come *SEO* o *Googlebomb* potrebbe essere più interessato a *PageRank*, che figura solo in una delle categorie dei termini elencati (ossia *Google*) ma è in categorie collegate ad esse e potrebbe quindi essere individuata con un'euristica che sfrutti a sua volta delle tecniche di assegnamento automatico come quelle che si vogliono esaminare. Infatti scegliendo come macrocategorie proprio le categorie che contengono le parole chiave o i tag già usati dall'utente si potrebbero individuare gli articoli a cavallo tra più macrocategorie, che avrebbero quindi una buona probabilità di avere a che fare con l'argomento effettivamente trattato dall'utente che non necessariamente corrisponde a una categoria specifica.

Si potrebbero suggerire ai visitatori articoli simili a quello che sta leggendo identificati grazie all'appartenenza di entrambi alle stesse macrocategorie scelte opportunamente. Allo stesso modo si potrebbero organizzare grandi quantità di testo dividendoli in base alle macrocategorie di appartenenza, una volta identificato un criterio per abbinare un testo a un articolo dell'enciclopedia.

Le tecniche discusse si potrebbero utilizzare con altre fonti di dati organizzate in maniera simile. Per esempio *DMOZ* è un progetto che mira a catalogare i siti web in base alle segnalazioni degli utenti, che li organizzano in categorie strettamente gerarchiche. Si potrebbero usare queste tecniche per catalogare il contenuto dei siti, considerando i link tra le loro pagine come delle appartenenze alle categorie. Tuttavia le categorie di DMOZ sono organizzate ad albero, quindi sarebbero necessarie delle modifiche anche sostanziali degli algoritmi.

Anche altri servizi, come il famoso sito di aste online eBay e numerosi siti di annunci, hanno una struttura a categorie che potrebbe prestarsi ad analisi di questo tipo.

Dopo avere scelto le *macrocategorie* da utilizzare è possibile effettuare nuovamente l'assegnamento automatico delle pagine alle macrocategorie ottenendo dei dati aggiornati alle ultime modifiche di Wikipedia e stabilire se alcuni criteri di assegnazione usati in passato da vari studiosi siano ancora validi. Le macrocategorie scelte dovrebbero rappresentare una sorta di partizione della conoscenza, quindi non dovrebbero esserci macrocategorie che rappresentano praticamente la stessa cosa o si contengono, come *Technologies* e *Applied sciences* o *History* e *Modern history,* nè argomenti che non sono coperti da nessuna macrocategoria. Infatti in tal caso non sarebbe possibile decidere con esattezza quale output dovrebbe dare l'algoritmo su un articolo comune a due macrocategorie o non classificabile in nessuna, e ci si troverebbe con dei dati privi di senso che costituirebbero un rumore nelle statistiche sulla struttura del progetto. Bisogna osservare però che non è un errore l'appartenenza di un articolo a più macrocategorie. Per esempio se si scegliessero come categorie *People*, *Geography*, *History*, *Nature*, *Science and technology*, *Sports* e *Humanities*, che rispettano le due condizioni indicate, la pagina *Garibaldi* sarebbe a cavallo tra *History* e *People*, così come la pagina *Jury Chechi* sarebbe contesa tra *Sports* e *People*. Questi non sono errori perché esistono sempre, qualunque sia la scelta delle macrocategorie, degli argomenti a cavallo tra due o più di esse, e anzi lo studio di quali siano le coppie di macrocategorie con la maggior quantità di articoli condivisi potrebbe essere molto utile per capire i legami tra i diversi argomenti.

Lo scopo della tesi non è solo quello di analizzare la struttura di Wikipedia, il modo in cui essa copre gli argomenti e come questi possono essere collegati tra di loro, ma anche e soprattutto cercare ed esaminare degli algoritmi che possano compiere questa operazione automaticamente su un set di macrocategorie qualsiasi tali da rispettare i due vincoli. Si sceglieranno quindi delle macrocategorie in un numero abbastanza alto per testare la robustezza delle tecniche in presenza di argomenti che potrebbero non es-

sere rappresentati molto chiaramente all'interno della struttura a categorie di Wikipedia. Infatti mentre argomenti come *Geography* sono rappresentati da una categoria omonima organizzata in maniera molto dettagliata in 25 sottocategorie come ad esempio *Geography terminology*, *Geocodes* e *Places* e quest'ultima contiene categorie come *Countries*, *Continents* e *Earth*, che contengono ognuna una struttura tassonomica, argomenti come *Agriculture* sono organizzati in maniera molto meno precisa. Le sottocategorie di *Agriculture* sono quasi tutte riconducibili ad altri argomenti, come *Agriculture by country* che è legata a *Geography*, *History of agriculture* che è legata ad *History* o *Agriculture minister*, che è legata a *Politics* (o a qualsiasi altra macrocategoria che contenga gli argomenti di politica).

Si noti che tutti gli algoritmi presentati riceveranno in input la lista delle macrocategorie, quindi si potrebbero replicare le procedure scegliendo delle macrocategorie diverse con una quantità minima di modifiche.

Capitolo 2

Stato dell'arte

Sono state svolte numerose ricerche su Wikipedia e sulla sua struttura a categorie (volte ad analizzare gli assegnamenti alle stesse generati dagli utenti) per effettuare delle analisi statistiche sia sull'andamento temporale del progetto che sul modo in cui l'enciclopedia copre i vari argomenti, cercando dei metodi per assegnare ogni articolo alla categoria più adatta tra quelle presenti in un insieme prestabilito. Numerose altre ricerche riguardano l'analisi del testo degli articoli dell'enciclopedia, per estrarne delle conoscenze da inserire nelle ontologie ricercando delle strutture grammaticali note all'interno dei testi e mappando gli articoli con i termini delle ontologie in modo da estrarre delle relazioni fra i termini mappati. Anche dal punto di vista dell'elaborazione del linguaggio naturale Wikipedia è stata usata in diversi lavori per estrarne delle informazioni come quelle contenute nel database WordNet, per estrarre iponimie, iperonimie e altre relazioni tra parole, così come dai collegamenti tra gli articoli nelle diverse lingue (chiamate *inter-Wiki*) è possibile estrarre informazioni utili per la creazione di traduttori automatici.

2.1 La struttura delle categorie di Wikipedia

Gli utenti di Wikipedia ne modificano i contenuti tramite la modifica del codice delle pagine e delle categorie. Questo codice è detto *WikiCode,* ed è riconosciuto dal software *MediaWiki* su cui si basa attualmente l'enciclopedia. L'utente può formattare il testo semplicemente aggiungendo degli apici prima e dopo le stringhe da formattare; tre apici indicano il grassetto, due il corsivo e cinque il corsivo e il grassetto.

Racchiudendo un termine tra parentesi quadre si crea un collegamento ipertestuale alla pagina su quel termine, con la possibilità di separare il

termine visualizzato da quello a cui porta il collegamento, in modo da poter creare collegamenti a una pagina utilizzandone sigle, diminutivi o omonimie che rendano più scorrevole la lettura di un articolo.

Le pagine di Wikipedia sono organizzate in *namespace*. Un namespace è una parola che precede il nome della pagina da cui è separata da due punti, in maniera uguale a quanto avviene con i namespace dell' XML o del C++. Vengono usati per distinguere i diversi tipi di pagina del progetto, ad esempio *Talk:SuperMario, User:SuperMario, Wikipedia:Help, Supermario* sono i nomi, rispettivamente, di una pagina di discussione su un articolo, di una pagina di un utente, di una pagina di aiuto o informazioni per gli editori stessi e di un articolo vero e proprio.

Esiste anche il namespace *Category,* che identifica le categorie. Una categoria ha infatti un suo contenuto, scritto in *WikiCode*, che appare come intestazione prima dell'elenco dei contenuti che invece è generato dal software *MediaWiki* a partire dalle indicazioni di appartenenza alle categorie inserite esplicitamente negli articoli stessi dagli editori tramite il codice *[[Category:nomecategoria]].*Ogni utente, registrato o meno, può quindi modificare il codice di una pagina e assegnarla a delle categorie arbitrarie (o rimuovere degli assegnamenti fatti da altri). Aggiungendo lo stesso codice all'interno dell'intestazione di una categoria, naturalmente, si assegna la stessa ad altre categorie, che devono essere pagine del namespace *Category.* In questo modo è impossibile assegnare una categoria o una pagina a un'altra pagina.

Esiste anche la possibilità di aggiungere alle pagine dei blocchi di codice di frequente utilizzo chiamati *template*; un template è una pagina nel namespace *Template* che viene richiamata scrivendone il nome tra doppie parentesi graffe accompagnato eventualmente da dei parametri che influenzeranno il comportamento del codice richiamato, in maniera simile alle funzioni dei linguaggi di programmazione. Una caratteristica notevole dei *template* è che possono assegnare la pagina che li richiama a delle categorie, il cui codice di assegnamento non appare esplicitamente all'interno della pagina stessa.Un esempio è la pagina *Milan*, di cui si riporta una sintesi:

```
{{Redirect|Milano}}
{{Other uses}}
'''Milan'''  is a [[city]] in [[Italy]] and the [[
   capital city|capital]] of the [[regions of Italy|
   region]] of [[Lombardy]] and of the [[province of
   Milan]].
[...]
==External links==
```

```
*[http://www.milano.it The main site of Milan]

 {{Regional capitals of Italy}}
{{Province of Milan}}
[[Category:Milan| ]] [[Category:Populated places
   established in the 1st millennium BC]]
```

I primi template, *redirect* e *other uses*, generano due avvisi della presenza di voci su argomenti ominimi, come la città di Milan in Texas o la squadra di calcio. Dopo il contenuto della voce, qui ovviamente omesso, vengono inseriti due template sinottici per potere vedere le voci sui capoluoghi di regione italiani e sulle voci sulla provincia di Milano. Quindi la pagina viene associata a due categorie, *Milan* e *Populated places established in the 1st millennium BC*. La scrittura con il simbolo — della categoria Milan serve a non inserire la pagina nell'elenco ma a specificare che è la pagina principale della categoria omonima. Tuttavia si osserva che Wikipedia associa la pagina anche alle categorie *Cities and towns in Lombardy* e *Communes of the Province of Milan*. Si tratta di due assegnamenti impliciti negli ultimi due template. Poiché un *template* può a sua volta richiamarne altri senza un limite teorico di annidamento, diventa molto difficile risalire alle categorie di assegnamento dalla semplice analisi del codice, ed è compito di *MediaWiki* rendere esplicite qeste informazioni. Si osserva anche la presenza di una sezione, identificata dalla sequenza ==, che verrà visualizzata con un titolo in grassetto, contenente dei collegamenti esterni. Questi ultimi vengono definiti tramite le singole parentesi quadre contenenti l'URL di destinazione e il testo cliccabile che verrà visualizzato.

2.2 Utilizzo delle categorie per l'inserimento automatico di conoscenza nelle ontologie

Sono stati svolti numerosi studi sull'utilizzo delle categorie di Wikipedia come base per la costruzione di ontologie e tassonomie. La costruzione di basi di conoscenza molto vaste e che coprano numerosi argomenti è infatti un tema cruciale nell'intelligenza artificiale, ma l'enorme difficoltà incontrata nella compilazione manuale di ontologie sufficientemente vaste e la rapidissima crescita di Wikipedia in temini di quantità di informazioni contenute ha portato molti a cercare delle tecniche per riversare le informazioni contenute in essa in ontologie utilizzabili dalle macchine.

Uno studio di Medelyan[16] ha mappato i termini del progetto *Cyc* con gli articoli di Wikipedia sfruttando la struttura a categorie. Il progetto cyc

mira a creare un'ontologia che copra praticamente ogni argomento della vita quotidiana per fornire alle intelligenze artificiali una base di *senso comune*, e lo studio evidenzia come delle fonti di Wikipedia contenga enormi quantità di informazioni inserite dagli utenti che possono essere usate per espandere l'ontologia cyc in maniera quasi automatica. Poiché anche nell'ontologia esistono delle categorie di concetti si possono utilizzare queste categorie per mappare con maggiore precisione i termini, sfruttando i mappaggi già eseguiti, assumendo che termini appartenenti alla stessa categoria di cyc tendano a corrispondere con articoli appartenenti alle stesse categorie di Wikipedia. Poiché i problemi delle omonimie e dei nomi diversi per indicare lo stesso concetto sono molto insidiosi, lo studio sfrutta le pagine di disambiguazione e di redirect per individuare termini con omonimi e con nomi diversi da quelli individuati. Una volta ottenute in questo modo le liste degli articoli che hanno la maggior probabilità di corrispondere a un termine, si sceglie quello più adatto osservandone le categorie di appartenenza e confrontandone il testo con quello degli altri articoli già abbinati a un termine.

Per la scelta dell'articolo più adatto tra quelli di una lista è stata elaborata, sempre da Medelyan, una tecnica[18] per utilizzare le pagine di Wikipedia come fonte di termini e di relazioni per il database Cyc. È stato quindi illustrato un metodo per abbinare un termine dell'ontologia al suo articolo corrispondente di Wikipedia, scegliendone tre con i nomi più simili al termine e analizzando quindi le similitudini tra i testi degli articoli simili, individuati grazie all'analisi della frequenza delle parole presenti, per decidere quale articolo dell'enciclopedia ha la massima similitudine con i tre proposti. Tale articolo viene poi analizzato alla ricerca di espressioni come *"X are a Y"* o *"The X is one of the Y"* per convertirle in proprietà dell'ontologia. È bene evidenziare come questo mappaggio tra l'ontologia e Wikipedia possa essere usato anche per tradurre dei concetti tra una lingua e un'altra grazie agli interwiki, per individuare dei sinonimi tra i nomi dei termini analizzando i redirect e gli abbinamenti tra il nome della pagina e il nome visualizzato definiti implicitamente nei collegamenti, ottenere delle descrizioni in linguaggio naturale dei concetti e indicizzare degli URL indicati tra i collegamenti esterni delle pagine di Wikipedia.

Il lavoro di Ponzetto[17] utilizza le categorie di Wikipedia per definire automaticamente delle tassonomie molto vaste. Vengono mostrati dei metodi per filtrare le categorie che non hanno un valore semantico ma sono usate dagli editori per organizzare le pagine con certe caratteristiche, per esempio quelle incomplete o quelle da fondere. Per eliminarle si esamina la categoria *Wikipedia_administration*, che le contiene, estraendone i contenuti. Tuttavia non tutto ciò che è contenuto in tale categoria è realmente privo di valore

semantico, quindi il processo deve essere osservato e controllato manualmente. Quindi, viene illustrato come molte categorie abbiano un nome del tipo "*X by Y*" e quindi non definiscano una sussunzione ma solo una ripartizione degli elementi volta a impedire che esistano categorie eccessivamente grosse, quindi il grafo viene modificato in modo che le categorie il cui nome segue tale schema siano integrate nella categoria che le contiene. Dopo aver effettuato queste manipolzioni dei dati è possibile usare le categorie per estrarre delle relazioni di tipo *isa*. Essendo queste delle relazioni transitive è possibile inferire numerose informazioni, per esempio se viene stabilito che *"MICROSOFT isa COMPANIES LISTED IN NASDAQ"* si inferisce automaticamente che "*MICROSOFT isa MULTINATIONAL COMPANIES*" sfruttando la struttura delle categorie.

Lo studio di Suchanek[13] mostra come sia possibile creare automaticamente un'ontologia reificata con oltre 15 milioni di fatti e quasi 2 milioni di termini tramite l'analisi di Wikipedia. Questa ontologia, chiamata *YAGO (Yet Another Great Ontology)*, fa uso sia dell'analisi euristica del testo degli articoli, alla ricerca di pattern del tipo "*X is a Y*", sia dell'analisi dei parametri passati ai template. Infatti esistono dei template che servono a visualizzare delle tabelle riassuntive in alcune pagine contenenti dei dati in un formato noto a priori. Per esempio la pagina su Elvis Presley contiene questo codice di template parametrico:

```
{{Infobox musical artist
|Name = Elvis Presley
|Birth_name = Elvis Aaron Presley
|Alias =
|Height = {{Height|feet=6}}
|Spouse = [[Priscilla Presley]] (1967-1973)
|Born = {{birth date|1935|1|8}}<br /><small>[[Tupelo,
   Mississippi]],<br />United States</small>
|Died = {{Death date and age|mf=yes
   |1977|08|16|1935|01|08}}<br /><small>[[Memphis,
   Tennessee]],<br />United States</small>
|Genre = [[Rock and roll]], [[pop music|pop]], [[
   rockabilly]], [[country music|country]], [[blues]],
    [[gospel music|gospel]], [[rhythm and blues|R&B]]
|Associated_acts = The Blue Moon Boys, [[The
   Jordanaires]], [[The Imperials]]
|Occupation = Musician, actor
|Instrument = Vocals, guitar, piano
```

```
|Years_active = 1954-1977
|Label = [[Sun Records|Sun]], [[RCA Records|RCA Victor
    ]]
|URL = [http://www.Elvis.com www.elvis.com]
|Notable_instruments= [[C. F. Martin & Company|Martin
    D-18]], [[Gibson J-200]] }}
```

È molto semplice creare un programma che estragga questi dati e li inserisca direttamente nell'ontologia, ottenendo grandi quantità di informazioni con una possibilità di errore di interpretazione quasi nulla. Le categorie non sono invece utilizzate da YAGO, poiché utilizza una tassonomia esterna chiamata *WordNet*.

Lo studio di Chernov[19] mostra delle tecniche per calcolare un coefficiente del livello di vicinanza semantica tra due categorie, basandosi sulla quantità di collegamenti tra le pagine appartenenti ai due insiemi. L'idea di base è che se una pagina, come *Country*, presenta numerosi collegamenti a un'altra, ad esempio *Capital*, esiste una correlazione semantica tra i due concetti rappresentati. Dunque, se le pagine contenute in una categoria presentano spesso dei legami con quelle contenute in un'altra è ragionevole supporre che le categorie stesse rappresentino dei concetti semanticamente vicini. Viene dunque calcolato il numero di collegamenti tra le pagine di due categorie e normalizzato in base al numero di pagine contenute in generale, in modo da non aumentare la correlazione categorie grosse. Lo studio ha mostrato che contando solo i link entranti nelle pagine si ottiene una buona similitudine tra le vicinanze indicate dai valutatori umani e quelle stabilite automaticamente dal programma.

2.3 Utilizzo delle categorie per Elaborazione del Linguaggio Naturale (NLP)

Poiché Wikipedia contiene del testo su ogni argomento, scritto da utenti diversi con diversi stili di scrittura, costituisce un prezioso banco di prova nello studio degli algoritmi per l'elaborazione del linguaggio naturale (*NLP, Natural Language Processing*). È infatti possibile utilizzare il testo degli articoli per produrre delle *WordNet*, ossia dei grafi di parole legate in base alle statistiche sulla tendenza ad apparire in un certo ordine e con una certa frequenza all'interno delle frasi. Queste statistiche eprmettono di dedurre con una certa approssimazione il ruolo della parola all'interno della frase e i suoi legami con altri termini con cui forma delle strutture idiomatiche.

Lo studio di Zesch[22] analizza il grafo degli articoli e delle categorie della

Wikipedia in lingua tedesca esteso con degli archi che indicano i collegamenti tra gli articoli e mostra che esistono delle similitudini statistiche tra il grafo delle categorie e le WordNet, come per esempio un simile diametro e una simile connettività. Grazie a queste similitudini è possibile applicare con successo degli algoritmi per le WordNet al grafo delle categorie e utilizzarlo come risorsa lessicale semantica. È necessario solo preprocessare il grafo per eliminare i cicli, operazione che verrà illustrata anche in questa tesi. Una volta eliminati diventa quindi possibile determinare se delle parole esprimono concetti che sono collegati tra di loro analizzando i legami tra i nodi del grafo delle categorie che rappresentano gli articoli omonimi, e spesso è possibile anche stabilire che tipo di legame sussiste.

2.4 Utilizzo del grafo delle categorie per stabilire l'argomento di un testo

Gli assegnamenti alle categorie inseriti arbitrariamente dagli editori possono essere usati per determinare l'argomento di un testo. Utilizzando delle tecniche come quelle allo stato dell'arte che verranno descritte fra poco e quelle discusse in questa tesi è possibile stabilire automaticamente a quale macrocategoria assegnare un articolo. Sfruttando questi assegnamenti automatici è possibile assegnare un testo a un certo argomento analizzandone le singole parole.

Lo studio di Syed[21] utilizza la cosine similarity per cercare gli articoli di Wikipedia più simili al testo da analizzare. Una volta individuati tali articoli, si assegnano dei valori probabilistici alle categorie che li contengono, che vengono sommati per stabilire empiricamente quali siano le categorie più adatte a contenere il testo di partenza. Un'altra variante analizzata comprende l'uso dei collegamenti tra le pagine come parametro aggiuntivo per stabilire i valori probabilistici, similmente a quanto fatto da Zesch.

2.5 Gli studi di Kittur e Holloway

Si vorrebbe trovare un algoritmo che effettui automaticamente l'assegnamento degli articoli alle macrocategorie scelte.

Uno studio di Kittur[15] ha mostrato che un metodo valido per svolgere questa operazione è calcolare le distanze topologiche tra i nodi rappresentanti le categorie scelte (che verranno chiamate da ora in poi *macrocategorie*) e quelli rappresentanti le pagine da analizzare all'interno del grafo di appartenenza di pagine e categorie all'interno dell'enciclopedia. Il percorso

più breve indica la macrocategoria a cui abbinare la pagina. Dei volontari reclutati tramite l'*Amazon Mechanical Turk Market* (*mturk.com*) hanno poi assegnato ad ogni pagina un punteggio del grado di appartenenza alle macrocategorie ripartendo un punteggio di 100 punti rappresentanti il grado di correlazione tra di esse. La correttezza di tale metodo automatico è stata così confrontata con quella dell'abbinamento effettuato manualmente dai volontari ottenendo una correlazione tra i risultati di 0.67. Lo studio ha anche mostrato che variando la tecnica, per esempio assegnando un valore di distanza normalizzato in base alla profondità tassonomica delle macrocategorie, non si ottengono risultati significativamente migliori.

La ricerca è stata svolta nel gennaio 2008, quando Wikipedia aveva 276834 categorie, circa 2 milioni di voci (ossia pagine contenenti effettivamente articoli dell'enciclopedia, non pagine di aiuto, discussione o comunque interne al progetto) e oltre 20 milioni di assegnamenti di pagine o categorie ad altre categorie, mentre a marzo 2010 si contano 3 milioni di pagine di contenuto, oltre 565108 categorie e oltre 40 milioni di assegnamenti.

Nello studio si è poi mostrato come questa tecnica possa essere applicata a diverse versioni del database di Wikipedia estratte nel corso del tempo per ottenere dei dati sull'evoluzione del progetto.

È stato così dimostrato, empiricamente, che dal luglio del 2006 al gennaio 2008 le voci sulla scienze naturali, sulla cultura e sulle arti sono più che triplicate, le voci sulla storia e sulla matematica sono raddoppiate e c'è stata una crescita del numero di voci su tutti gli argomenti a parte la tecnologia, scesa del 6%.

Kittur non fornisce spiegazioni sul motivo della diminuzione, che potrebbe essere causata da una diminuzione del numero delle pagine o, più probabilmente, da una diversa struttura delle categorie che coprono gli argomenti tecnologici che ha portato ad assegnarne le pagine ad altre aree semantiche.

Un'altra possibilità derivante dall'assegnamento delle pagine alle macrocategorie stabilite è lo studio dei conflitti di delle stesse. Un conflitto si ha quando un testo viene modificato da un utente e poi riportato alla forma precedente da un altro utente, più volte. In genere questo evento si verifica nel caso di un vandalismo, ossia una modifica palesemente sbagliata che viene annullata dagli utenti che la individuano, o nel caso di diverse opinioni sul contenuto, caso che viene gestito in genere annullando la modifica ritenuta sbagliata e discutendone nella pagina di discussione abbinata all'articolo. Questi eventi si possono rilevare automaticamente cercando le modifiche che annullano l'effetto di modifiche precedenti, e Kittur ha dimostrato come gli argomenti su cui si verificano più conflitti sono *Religion* e *Philosophy*, entrambi oggetto dle 28% dei casi, seguiti da *People*, al 14% e

Science, al 10%.

Un altro lavoro[20] ha analizzato la struttura a categorie di Wikipedia per estrarne le proprietà statistiche, confrontandone alcune con quelle dell'*Enciclopedia Britannica* e di *Microsoft Encarta.*

Inoltre, il lavoro illustra graficamente la distribuzione di frequenza della dimensione delle categorie in termini di articoli contenuti direttamente e, viceversa, del numero di categorie abbinate ad ogni articolo.

Gli autori hanno creato una mappa visuale, a due dimensioni, delle pagine e delle categorie di Wikipedia nel 2006, dove la vicinanza tra gli articoli è data dal numero di collegamenti presenti nelle pagine, mentre la vicinanza delle categorie agli articoli o alle altre categorie è frutto della comparazione effettuata con la tecnica della cosine similarity.

Questo metodo, che verrà ampiamente discusso e usato in questa tesi, ha portato a considerare graficamente vicine le categorie che hanno molto articoli in comune.

Dopo aver tracciato questa mappa gli autori hanno evidenziato i punti corrispondenti alle pagine create in vari periodi, mostrando come la mappa campi nel corso del tempo e evidenziando che le categorie sono oggetto di frequenti modifiche, se si escludono quelle create automaticamente dai bot.

Colorando invece i punti della rappresentazione corrispondenti alle pagine o alle categorie la cui ultima modifica al momento dell'estrazione dei dati era stata effettuata dallo stesso utente, selezionando solo i 7 utenti più attivi per avere dei dati sufficienti, si osserva che i punti dello stesso colore, ossia relativi allo stesso utente, formano delle macchie abbastanza compatte.

Questo è un altro dato a favore dell'efficacia degli assegnamenti tra categorie e tra pagine nello stabilire la vicinanza semantica degli argomenti. Infatti si presume che ogni utente abbia degli interessi e delle aree di competenza precise e tenda, quindi, a modificare o creare pagine su tali argomenti. Se le modifiche apportate sono rappresentate da macchie di colore compatte significa che i punti corrispondenti a pagine su argomenti semanticamente correlati sono stati posizionati vicini nella rappresentazione grafica, e quindi sono anche topologicamente vicini.

A ulteriore conferma di ciò, si nota che le modifiche automatiche effettuate dai 3 bot più attivi sono sparse su tutta la mappa casualmente, a parte uno, *Rambot.* Infatti questi bot non si concentrano su un argomento specifico (a parte Rambot) ma intervengono su certe strutture la cui presenza è indipendente dal contenuto della pagina.

I tre bot si chiamano *Whobot*, *KocjoBot* e *Rambot.* Nello specifico il bot *Whobot* si occupa di annotare le pagine che potrebbero essere da rivedere o da cancellare inserendovi una nota e aggiornare gli *interwiki* (i collegamenti

alla stessa voce in altre lingue), *KocjoBot* di aggiornare gli *interwiki* tra le pagine mentre *Rambot* si occupa delle città americane e degli *interwiki*. *Rambot* produce delle modifiche che appaiono come delle piccole macchie e non come punti totalmente sparsi come nel caso degli altri due bot, questo è spiegato dal fatto che agisce sulle pagine riguardanti un certo argomento, ossia le città degli Stati uniti.

Un applicazione pratica di queste tecniche si trova nel lavoro di Cosley[12], che riguarda la creazione di un bot, *SuggestBot*, che consiglia agli utenti quali articoli potrebbero migliorare.

Nelle comunità virtuali, infatti, il problema della suddivisione del lavoro è molto difficile da affrontare e raramente si suggeriscono agli utenti i lavori da compiere se non con criteri casuali.

SuggestBot, tuttora attivo e raggiungibile alla sua pagina[7] di Wikipedia utilizza il coefficiente di Jaccard[3], che è equivalente alla cosine similarity nel caso di attributi binari, per confrontare la lista delle pagine modificate da un utente con le liste delle pagine modificate dagli altri utenti partendo dal presupposto che se due utenti hanno molte pagine in comune tra quelle modificate ognuno sarà interessato a migliorare gli articoli modificati dall'altro.

Una volta individuati questi articoli *SuggestBot* li suggerisce all'utente perché possa avere delle indicazioni sul lavoro da compiere.

Il lavoro di Gabrilovich[14] ha utilizzato i dati contenuti in Wikipedia per determinare quali sono i concetti chiave in un testo.

Analizzando il testo degli articoli dell'enciclopedia e utilizzando la cosine similarity per confrontarli con il blocco di testo da analizzare, il lavoro ha mostrato che le voci con il maggior coefficiente di similitudine con il testo da analizzare sono generalmente quelle corrispondenti ai concetti chiave dell'articolo, mostrando la validità della cosine similarity come criterio di valutazione.

Questo risultato fa pensare che, se si riuscisse a stabilire automaticamente a che macrocategoria assegnare un certo articolo, si potrebbero assegnare dei testi alle macrocategorie esaminandone il contenuto.

Capitolo 3

Obiettivi e metodologia

L'obiettivo che ci si prefigge in questa tesi è dunque quello di valutare l'efficacia dell'algoritmo di Kittur nell'effettuare l'assegnamento con una selezione più ampia di macrocategorie e su una quantità di dati maggiore in termini di numero di pagine, di categorie e di relazioni di assegnamento. Si vuole, in sostanza, valutare se il criterio della distanza topologica è valido anche in questo contesto più caotico.

Inoltre, si vogliono provare dei metodi alternativi, come la normalizzazione in base alla distanza tassonomica, l'assegnamento di costi di attraverso differenziato agli archi in base a certe proprietà come l'orientamento o la quantità di categorie contenute nei nodi di partenza e studiare degli algoritmi più complessi che valutino i percorsi possibili nel loro insieme in modo da cercare di diminuire l'effetto deleterio di alcune categorie e relazioni di appartenenza che avvicinano dei nodi poco correlati semanticamente. Ci si propone di effettuare gli assegnamenti secondo le diverse varianti dell'algoritmo basato sulla distanza topologica o in modi diversi per poi valutare la qualità dei risultati ottenuti automaticamente tramite il confronto con gli assegnamenti effettuati da un essere umano. In questo modo sarà possibile stabilire se esistono criteri di assegnamento che mostrano maggiore precisione della distanza topologica e che quindi hanno migliori possibilità di essere efficaci nell'elaborare dati privi di una struttura regolare e omogenea.

Le relazioni di appartenenza degli articoli e delle categorie stesse ad altre categorie portano intuitivamente a rappresentare i contenuti nella forma di un grafo con archi orientati.

Le relazioni tra i nodi di questo grafo potrebbero essere di un solo tipo generico chiamato appartenenza, ma si è preferito crearne due tipi: una che rappresenti l'appartenenza di una categoria a un'altra, che chiameremo SUBCATEGORYOF, e una che rappresenti l'appartenenza di una pagina a

una categoria, che chiameremo BELONGSTO.

Questa differenziazione è utile se non necessaria perché le due relazioni hanno delle proprietà diverse che si vorrebbero poter riconoscere facilmente, in particolare le relazioni di appartenenza di una categoria possono essere entranti o uscenti, mentre quelle di appartenenza di un articolo a una categoria possono solo essere uscenti dall'articolo poiché un articolo non può essere assegnato a un altro articolo ma al più a una categoria omonima.

Il grafo dunque sarà multi-relazionale, come rappresentato nella Figura3.1, e le categorie non rappresentano affatto una tassonomia, nonostante in genere seguano comunque una struttura gerarchica, ma vengono usate più come dei tag, per indicare al visitatore altre pagine o categorie a cui può essere interessato.

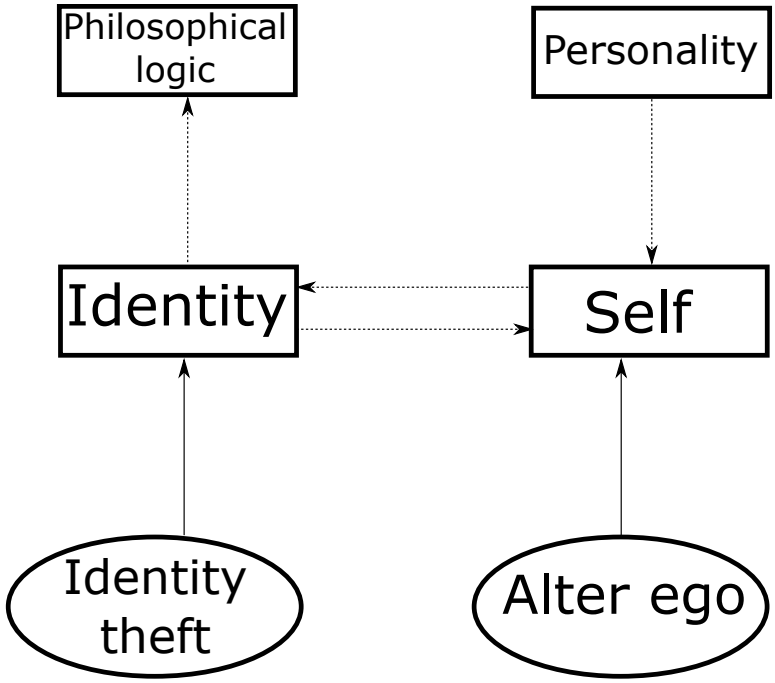

Figura 3.1: Una rappresentazione del grafo degli assegnamenti. Gli elementi negli ovali sono le pagine, nei rettangoli le categorie. Le frecce continue rappresentano l'appartenenza di una pagina a una categoria, quella tratteggiate l'appartenenza di una categoria a un'altra.
Gli elementi rappresentati sono realmente presenti nel dump di en.wikipedia trattato. Si noti che le categorie *Self* e *Identity* si contengono reciprocamente, formando un anello.

Possono addirittura esserci dei cicli di categorie che si contengono re-

ciprocamente, categorie che si auto-contengono, categorie vuote o non contenute in nessun'altra.

3.1 Il criterio di assegnamento di Kittur: la distanza topologica dalle macrocategorie

L'idea alla base dello studio di Kittur e anche di quello di Holloway è che le relazioni di appartenenza tra le categorie, ossia le relazioni *SUBCATEGORYOF*, siano relazioni di similitudine semantica, ossia colleghino categorie su argomenti che sono generalmente considerati legati fra di loro.

Dunque, più archi è necessario percorrere nel grafo delle appartenenze per collegare due nodi più i due nodi si possono considerare semanticamente distanti, ossia legati ad argomenti tra i quali non è evidente nessun collegamento.

Tuttavia, la struttura del grafo non è per niente omogenea, quindi si hanno argomenti coperti da categorie organizzate in maniera strettamente gerarchica e che quindi richiedono molti passaggi per essere attraversate e argomenti coperti, al contrario, da una rete di categorie fortemente interconnesse che si può attraversare con un numero esiguo di passaggi.

Inoltre, la vicinanza semantica non è una relazione transitiva, quindi due categorie A e B possono essere collegate a una categoria comune C ma rappresentare argomenti totalmente diversi. Ad esempio la pagina *RAI* appartiene a *Orphan initialisms* (categoria che contiene le pagine riguardanti organizzazioni o cose che vengono ancora chiamate con un nome vecchio e divenuto ora privo di significato) così come *Laser*, ma non è per niente intuitivo il legame fra l'azienda televisiva e la tecnologia per emettere luce monocromatica e concentrata.

Inoltre, Wikipedia contiene anche numerose categorie ad uso interno del progetto, come *Wikipedia categories in need of attention*, che elenca le categorie troppo grandi, troppo piccole, oggetto di frequenti vandalismi o con altre caratteristiche che richiedono un intervento da parte degli utenti. Un altro caso fra i tanti è la presenza di categorie che raggruppano gli articoli in base alla qualità e alle dimensioni. Naturalmente queste categorie sono deleterie ai fini del calcolo dell'appartenenza perché non hanno valore semantico (Figura 3.2), quindi sarà necessario eliminarle.

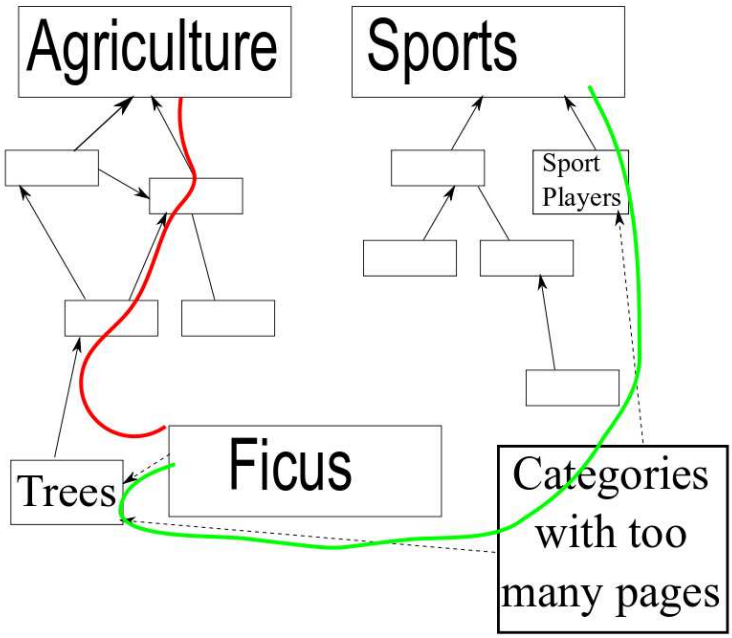

Figura 3.2:
Una categoria senza valore semantico collegata alle altre categorie altera il calcolo dei percorsi minimi rendendo vicini dei nodi che non hanno legami semantici. In questo caso un'ipotetica categoria che raccoglie le categorie con troppe pagine lega *Trees* e *Sport players* alterando il calcolo delle distanze dalle macrocategorie *Agriculture* e *Sports* e assegnando la voce *Ficus* alla seconda.

3.2 Scelta delle macrocategorie

Si vogliono scegliere delle macrocategorie che, insieme, coprano qualsiasi contenuto dell'enciclopedia. Inoltre gli argomenti scelti devono essere tali da non rendere difficile a un valutatore umano decidere a quale macrocategoria (o quali nel caso di voci che toccano degli ambiti interdisciplinari, come nel caso dell'epistemologia) assegnare una certa voce. Quindi si devono scegliere delle macrocategorie che, approssimativamente, generino una partizione della conoscenza: non si devono sovrapporre troppo spesso ma non devono nemmeno lasciare scoperti alcuni articoli.

Nel caso di Kittur le macrocategorie scelte erano 11, ora ne verranno usate 21.

Infatti, poiché le dimensioni di Wikipedia sono quasi raddoppiate, sia come numero di articoli che di categorie, sarebbe interessante determinare

se la tecnica della distanza topologica funziona anche in questo caso o se è necessario apportare delle modifiche.

Le macrocategorie sono state scelte prendendo spunto dalle categorie elencate in *main_topic_classifications* [4] e da quelle precedentemente usate da Kittur.

Delle 23 categorie in *main_topic_classifications* alcune non sono state selezionate perché hanno un contenuto difficilmente definibile, come *Life* o *Nature*, altre perché sono troppo generiche, come *Humanities* che potrebbe essere scomposta in *Culture* e *Arts*. Altre categorie hanno un significato molto simile, come *Technology* e *Applied sciences*, e quindi sono state fuse.

La categoria *History* ha due categorie che hanno un ruolo simile, ossia *Events* e *Chronology*.

Esiste infine una categoria di nome *Places* che pur non rientrando in queste 23 ha un ruolo simile a *Geography*.

Si potrebbero unire queste categorie dal significato molto simile in tre singole macrocategorie create ad hoc.

Un risultato interessante è la misura statistica della quantità di assegnamenti per articolo, ossia la quantità di articoli assegnati a un certo numero di macrocategorie, per capire quanto è frequente che un articolo venga assegnato a una sola o che sia invece assegnato a molte categorie contemporaneamente.

3.3 Possibili criteri alternativi alla distanza topologica

Si vedrà che che l'assegnamento degli articoli alle macrocategorie basato sulla semplice distanza topologica, che chiameremo *caso base*, può sbagliare principalmente per tre motivi:

- Esistono delle categorie prive di valore semantico, ad uso interno del progetto.
- Il percorso tra una categoria e una macrocategoria segue le relazioni SUBCATEGORYOF in entrambe le direzioni.
- Ogni argomento presenta una certa profondità tassonomica, quindi se per giungere a Biology dalla pagina di una specie animale bisogna risalire per varie categorie che rappresentano una tassonomia di famiglie, ordini e regni animali, tali che ognuna contenga poche sotto-categorie, per giungere ad Agriculture bisogna in genere passare per poche cat-

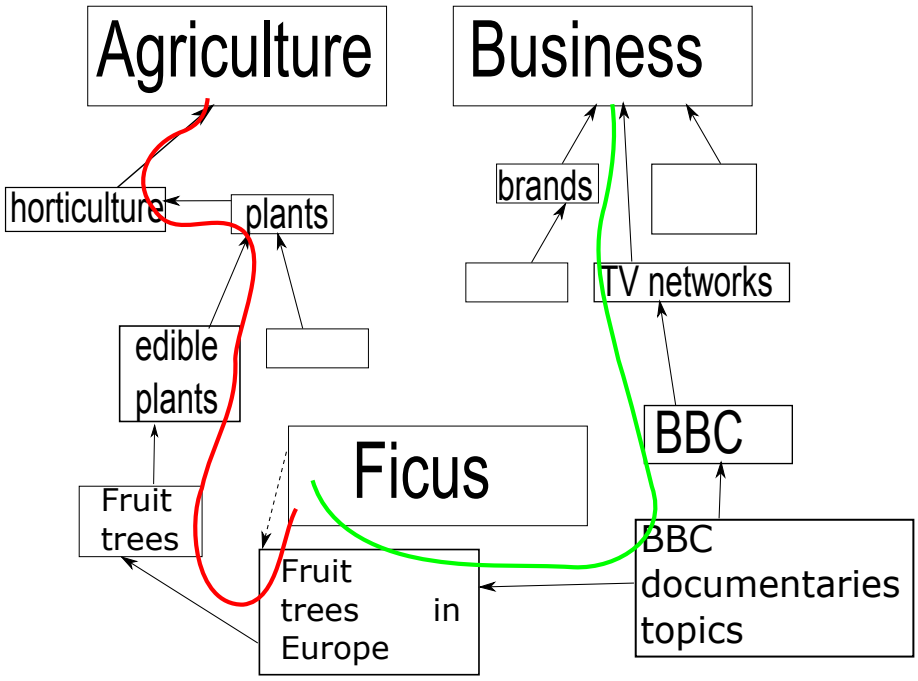

Figura 3.3:
La presenza di categorie a breve distanza da una macrocategoria, in questo caso *Fruit trees in Europe,* può causare l'assegnamento a quella macrocategoria delle categorie contenute con essa in un'altra, in questo caso *Fruit trees*

egorie molto grosse e generiche. Questo fa sì che Agriculture riceva molti più assegnamenti di quelli che le spetterebbero.

Il primo problema sarà risolto individuando e eliminando a mano questa categorie.

Il secondo problema è costituito dalla possibilità, durante la ricerca del percorso minimo, di muoversi lungo la relazione *SUBCATEGORYOF* anche in verso contrario, creando spesso delle anomalie dovute alla grande vicinanza di una categoria, contenuta nella stessa categoria di quella che si sta esaminando, a una macrocategoria, come mostrato nella Figura 3.3.

Infatti, se una categoria A appartiene a una categoria B che contiene pure la categoria A', e A' è molto vicina a una certa macrocategoria M, allora A, così come le altre categorie contenute in B, tenderà ad essere assegnata a M invece che alla macrocategoria a cui appartiene B e a cui generalmente verrebbe assegnata da un valutatore umano.

Le soluzioni più semplici a questo problema sono due:

- Calcolare i percorsi minimi percorrendo il grafo solo nella direzione

delle relazioni ogni volta che è possibile. Se non è possibile, allora si considera la macrocategoria a distanza infinita dal nodo in esame.

- Utilizzare una metrica per il calcolo dei costi che assegni un costo maggiore ai passaggi effettuati in direzione contraria alle relazioni *SUBCATEGORYOF*.

Il terzo problema è il più complesso, poiché è evidente dalle statistiche mostrate prima quanto frequentemente possa alterare gli assegnamenti, ed il più insidioso, in quanto la differente profondità tassonomica è un problema intrinseco del grafo degli assegnamenti che non è risolvibile a mano semplicemente eliminando o modificando un piccolo gruppo di categorie.

Una possibile tecnica è moltiplicare tutte le distanze topologiche di ogni categoria da ogni macrocategoria per un coefficiente tale da normalizzare i baricentri delle curve di distribuzione di frequenza delle distanze.

Si potrebbe anche combinare questa tecnica con l'assegnamento tramite percorsi che seguono l'orientamento degli archi, normalizzando le distanze topologiche relative a questo vincolo.

Un altro metodo per affrontare il problema potrebbe consistere nell'assegnare dinamicamente il costo di attraversamento di un arco basandosi sulle proprietà locali del grafo, come il numero di articoli o di categorie contenuti nelle due categorie legate dall'arco, la presenza di sotto-stringhe particolari nei nomi e molti altri fattori. Quest'ultima possibilità, essendo la più complessa per la quantità enorme di variazioni che si possono applicare, viene lasciata per ultima in modo da poter sfruttare i dati ricavati precedentemente dalle altre.

Infine, esiste la possibilità di tenere conto non solo del percorso minimo lungo il grafo, ma anche della quantità di percorsi distinti o con nodi in comune con cui è possibile raggiungere una certa pagina a partire da una macrocategoria.

L'idea di base è che una singola categoria può avere degli archi entranti o uscenti che collegano due categorie su argomenti che altrimenti sarebbero molto lontani in termini di distanze topologiche nel grafo, generando delle anomalie.

Se però si considerano più percorsi possibili contemporaneamente l'effetto di questi collegamenti viene ridotto, poiché per loro natura sono casi isolati mentre la maggior parte dei percorsi esistenti porterà alla macrocategoria corretta (Figura 3.4).

Si possono dunque assegnare alle categorie dei valori di appartenenza alle macrocategorie che tengano conto dei diversi modi in cui è possibile raggiungerle.

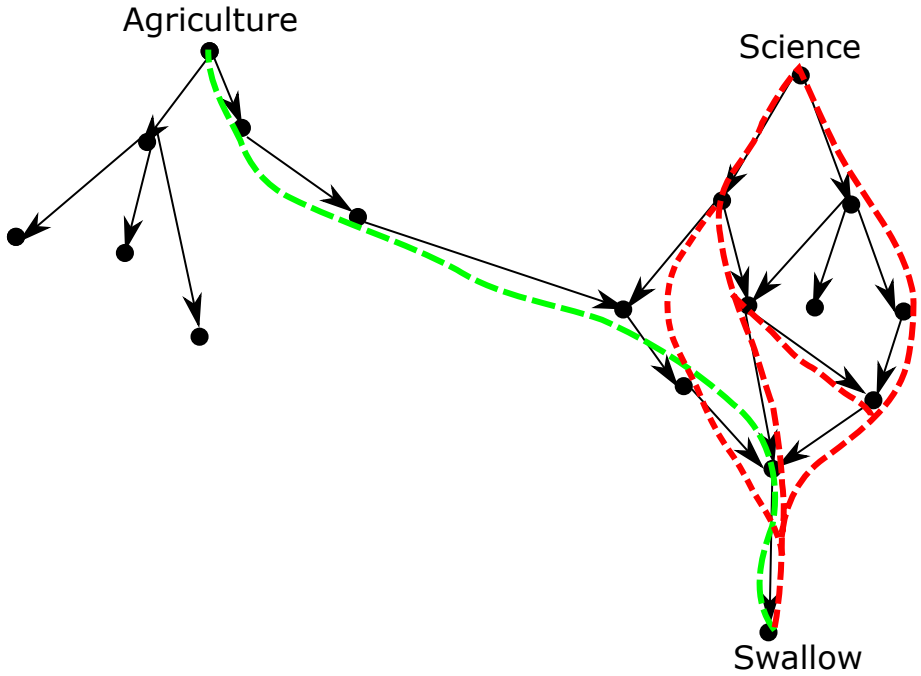

Figura 3.4:
Un arco solo può influenzare la distanza topologica tra una categoria e una macrocategoria e quindi alterare l'assegnamento degli articoli alle macrocategorie. Può dunque essere conveniente considerare i diversi percorsi possibili dalle macrocategorie nel loro insieme.

Capitolo 4

Progetto e realizzazione

Il lavoro si articola in varie fasi, illustrate in Figura 4.1, da svolgersi necessariamente nel giusto ordine. In alcuni casi esistono più varianti che si vogliono valutare, quindi vengono usati più volte i dati prodotti dalle fasi precedenti per delle elaborazioni diverse. Per esempio i vari algoritmi di calcolo delle distanze ricevono in input il grafo filtrato delle categorie senza valore semantico, che essendo lo stesso per tutti non viene naturalmente generato dal database ogni volta.

La normalizzazione delle distanze tassonomiche è effettuata basandosi sui risultati di un assegnamento, che vengono usati per calcolare i baricentri delle curve di distribuzione delle frequenze delle distanze dalle macrocategorie, dunque bisogna effettuare un assegnamento preliminare per poter calcolare i parametri di normalizzazione e quindi effettuare nuovamente l'assegnamento dopo aver normalizzato le distanze.

4.1 Creazione e filtraggio del grafo

Il dump di Wikipedia è disponibile in formato XML o MySQL. Si è utilizzato quest'ultimo perchè estrarne solo le informazioni sui nomi delle pagine e delle categorie e sulle loro relazioni, ignorandone il contenuto, è un'operazione immediata. Inoltre è più semplice iterare sulle tuple di una tabella per estrarne tutto il contenuto che effettuare la stessa operazione su un file XML, nonostante anche quest'ultimo caso sia tipico e semplice da affrontare.

Per svolgere il lavoro è stato generato un grafo delle categorie e delle pagine di Wikipedia, con le relative relazioni di appartenenza, memorizzandolo in Neo4j[6], un database non-relazionale.

Si è preferito un database non relazionale e basato sui grafi a un database relazionale perché, essendo il modello un grafo vero e proprio, i dati vengono

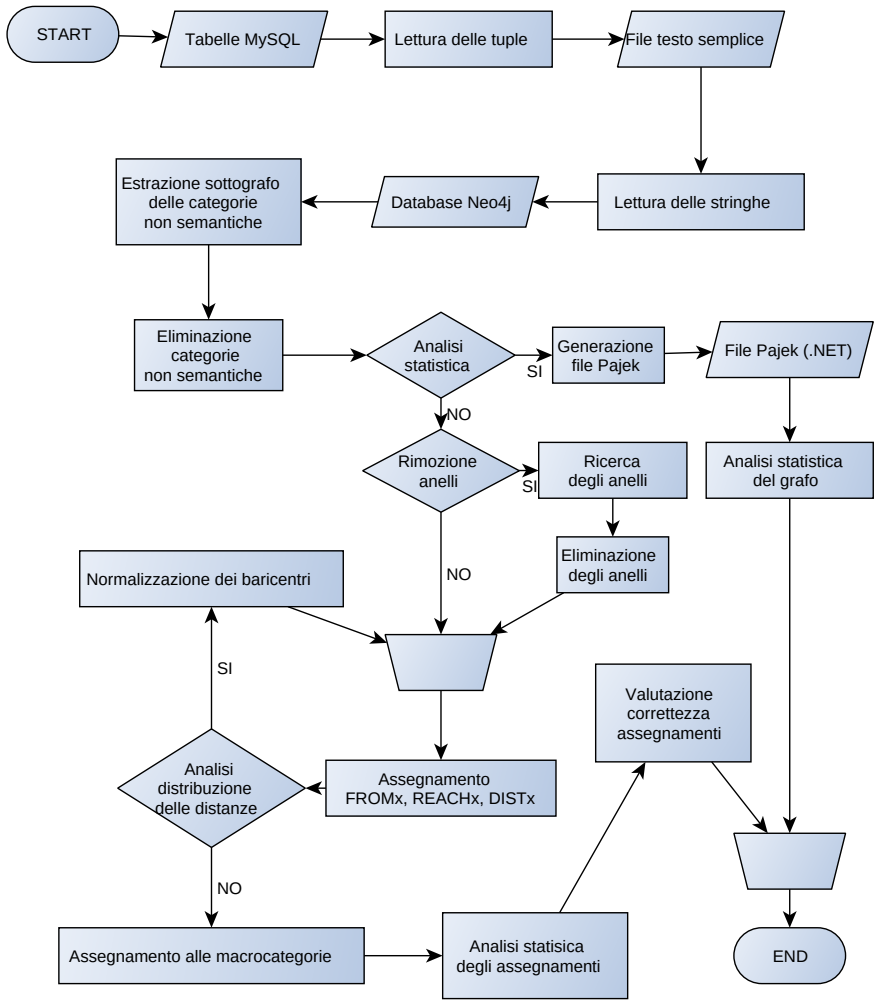

Figura 4.1:
Le fasi del lavoro, a partire dai dump in formato MySQL. Gli elementi rettangolari indicano le fasi di elaborazione dei dati, i parallelepipedi le fasi i I/O. Le frecce indicano consequenzialità logica e temporale, alcuni elementi formano, insieme ai blocchi di decisione, degli anello ad indicare che i risultati ottenuti servono a svolgere nuovamente un' operazione con dei parametri di funzionamento perfezionati proprio in base ai nuovi risultati.

gestiti con maggiore efficienza. In un database relazionale per definire le relazioni molti a molti, come quelle tra le categorie, ci vorrebbe una tabella apposita da leggere tramite più letture consecutive, eventualmente implicite in un join.

In Neo4j ogni nodo contiene al suo interno l'indice dei nodi collegati, velocizzando molto la ricerca che non è appesantita dalla grande quantità di nodi come accadrebbe se ci fosse un unico indice degli archi[5]. Il risultato è che se si esplora il grafo muovendosi di nodo in nodo non si notano differenze di velocità tra un grafo con milioni di nodi e uno con poche centinaia, dato che la ricerca dei nodi collegati a uno dato è basata su un indice dedicato e non su un indice che contiene tutte le relazioni. Lo svantaggio è che le operazioni globali, come la ricerca di relazioni con una certa caratteristica su tutto il grafo, diventano più lente.

Questo database permette di rappresentare un grafo manipolandone nodi e archi e assegnandogli delle proprietà, che possono essere valori numerici, stringhe o booleani, identificate in base al nome, le quali possono essere recuperate e lette in un secondo momento. Gli archi hanno anche un tipo, in questo caso la relazione di appartenenza di una categoria a un altra è rappresentata con un arco *SUBCATEGORYOF*, mentre quella di un articolo a una categoria è rappresentata con un arco *BELONGSTO*. Sia le pagine che gli articoli sono rappresentate da nodi che possiedono la proprietà *name*, di tipo stringa e che contiene il nome dell'elemento. Inoltre i nodi hanno la proprietà *idpag* oppure *idcat* contenente il numero identificativo della pagina o della categoria.

Per individuare e eliminare le categorie prive di valore semantico si è estratto ricorsivamente il sotto-grafo delle categorie contenute in *Wikipedia administration* (Figura *4.2*), come fatto nello studio di Ponzetto*[17]*.

Il procedimento è stato svolto in maniera semi-automatica, utilizzando delle correzioni manuali per evitare di estrarne delle sotto-categorie in realtà dotate di valore semantico, come quelle contenute in *Categories for renaming*: in questi casi bisogna impostare manualmente il programma per non estrarre il loro contenuto durante la ricorsione. Una volta ottenuto il sotto-grafo si itera sui suoi nodi e si cercano tali nodi nel grafo originale, in modo da eliminarli e rimanere solo con le categorie significative.

Altri due casi che devono essere gestiti dal software sono quelli delle pagine di *redirect* e *disambigua* (*disambiguazione*).

Una *disambigua* è una pagina che contiene un elenco di link a pagine che trattano vari argomenti con lo stesso nome. Ad esempio alla pagina *Java* corrisponde la voce sull'isola di Java in Indonesia, ma nella pagina *Java_(*disambiguation*)* possiamo vedere oltre una decina di link alle voci sui

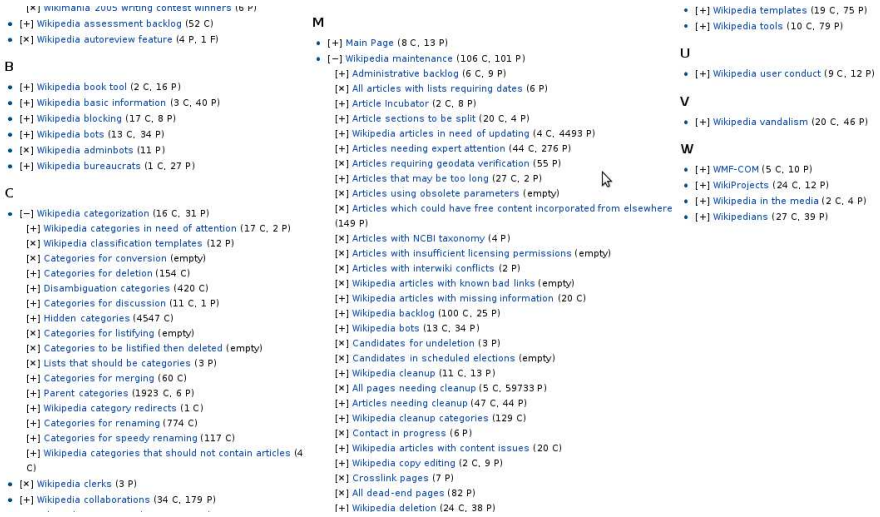

Figura 4.2: La visualizzazione via web della categoria *Wikipedia administration*

vari significati del termine: il linguaggio di programmazione, una marca di sigarette russe, un tipo di caffè, un pipistrello, un tipo di pollo, un distretto della Georgia e altre cose. Esse vengono identificate perché appartengono a categorie che contengono sempre la sotto-stringa "*isambiguation*" nel nome, quindi eliminate. Esistono circa 50,000 disambigue.

Un *redirect* è una pagina che reindirizza automaticamente i visitatori su un'altra pagina con il nome corretto. Per esempio "*George W.Bush*" redireziona a "*George W. Bush*", senza lo spazio prima del cognome. I *redirect* sono numerosissimi, anche perché vengono creati automaticamente dal software *MediaWiki* quando si sposta una pagina allo scopo di mantenere i link alla vecchia versione funzionanti. Si possono individuare e rimuovere perché appartengono sempre a categorie il cui nome contiene la sotto-stringa "*edirects*". Esistono circa 4 milioni di *redirect*s.

Si è deciso che alcune categorie già esistenti, come *Technology* e *Applied sciences*, *Geography* e *Places o History, Events e Chronology* sono da considerarsi come parte della stessa macrocategoria. Un altro caso è la categoria *Chronology*, che contiene delle categorie che raggruppano i millenni, le ere, i secoli, gli anni, i mesi e i giorni. Contiene inoltre ciò che ha a che fare con il tempo, come gli anniversari, i metodi di datazione storica, i calendari, le date in altre culture e le categorie che organizzano gli elementi in base al tempo, come *People by time*.

Esiste infine la categoria *Events* che contiene gli eventi organizzati sia per tipo, come *Accidents,* che per luogo, come *Events in Africa*, che per tipo e tempo, come *Conflicts by year*.

Si crea dunque un programma che inserisca nel grafo tre nuove categorie, chiamate *Geography and places*, *History and events* e *Technology and applied sciences*, a cui vengono assegnati dei numeri seriali arbitrari diversi tra di loro e diversi da quelli già usati. In ognuna di queste vengono poi inserite le due categorie indicate dai nomi, in più alla categoria sulla storia vengono aggiunte anche alcune di quelle contenute in *Chronology*, scelte a mano per escludere gli elementi, come la categoria *Months,* che non hanno a che fare con la storia. Dalla categoria *Events* vengono infine eliminate le categorie che raggruppano gli eventi per luogo o per tipo.

4.2 Analisi delle caratteristiche globali del grafo

Una volta generato e riadattato il grafo come illustrato è possibile esportarlo nel formato utilizzato da Pajek[9], un popolare tool di elaborazione dei grafi, e analizzarlo tramite il pacchetto igraph (versione 0.5.3)[2] di R (versione 2.6.2)[8].

Si è scelto Pajek[9] per la sua buona capacità di rappresentare grafi molto grossi, per l'abbondanza di strumenti di analisi matematica molto completi compatibili con questo formato e per la sua semplicità, che rende veloce l'esportazione da Neo4j.

Un file in Pajek è infatti un semplice file di testo con estensione *.NET* con la seguente struttura:

```
*Vertices 3
1 "articolo1"
2 "categoria1"
3 "categoria3"
*Arcs
1 2 1
```

La prima riga segnala l'inizio della dichiarazione dei vertici, di cui riporta il numero. Segue l'elenco dei nodi identificati con un numero seriale e una lista di proprietà (il numero identificativo e il nome della pagina o della categoria). Infine, se necessario, si dichiara l'inizio dell'elenco degli archi orientati e li si scrive, uno per riga, mettendo nell'ordine i numeri identificativi del nodo di partenza e di arrivo separati da uno spazio e, se necessario, il peso degli archi, che in questo caso, trattandosi di una rete non pesata, non viene specificato.

È poi possibile definire degli archi non orientati, con una terza sezione dalla sintassi identica a quella usata per gli archi ma con intestazione *Edges*, che non verrà usata perché si usano solo archi orientati.

È da notare il fatto che nonostante il grafo che si sta trattando abbia due tipi di archi, ossia *BELONGSTO* e *SUBCATEGORYOF*, Pajek non permette di definire la differenza e considera tutti gli archi dello stesso tipo. Questo però non è un problema, dato che il calcolo della distanza e tutti i risultati ottenuti non ne risentono.

Questo file può essere analizzato con vari strumenti matematici tra i quali si è scelto igraph[2], che serve proprio ad analizzare i grafi. Igraph è un pacchetto di R[8], un linguaggio di programmazione sviluppato per l'analisi matematica e statistica e abbinato a un omonimo IDE. È stato scelto perché è open source, efficiente, ricco di funzioni e abbastanza semplice da usare. Altre analisi possono essere condotte con degli programmi creati apposta per i compiti specifici, come contare il numero di pagine assegnate a un certo numero di categorie, calcolare il numero medio di categorie per pagina o cercare la pagina con più assegnamenti esistenti.

4.3 Assegnamento alle macrocategorie con il criterio di Kittur

Utilizzando l'algoritmo di Djikstra per calcolare una a una le distanze dei nodi pagina dai 21 nodi rappresentanti le macrocategorie si osserva che il tempo necessario a effettuare il calcolo è eccessivamente alto. Servirebbero infatti 63 milioni di ricerche di percorso minimo e, osservando che ognuna dura vari secondi, a volte anche una decina, si stima un tempo di elaborazione nell'ordine degli anni.

Si può ottimizzare il calcolo osservando che un percorso minimo tra due nodi contiene anche i sotto-percorsi minimi tra i nodi intermedi e gli estremi. Dunque ogni volta che si calcola una distanza tra un nodo categoria e una macrocategoria si assegna a ogni nodo intermedio una proprietà chiamata *FROMx* dove x è il nome della macrocategoria. Questa proprietà sarà un numero intero che indicherà la distanza topologica della categoria dalla macrocategoria, quindi, ad esempio, se una categoria dista 12 da *Geography and places* il suo nodo avrò la proprietà FROMGeography posta a 12. Se tale proprietà risulta già assegnata si evita di calcolare il percorso minimo e si restituisce direttamente il valore della proprietà.

L'algoritmo è:

- Itera sulle categorie che contengono la pagina (ossia i nodi che hanno un arco BELONGSTO con un nodo che ha la proprietà idcat)
  - Quando una di esse possiede la proprietà FROMx, dove x è il

nome della macrocategoria, utilizza quel valore come distanza della categoria dalla macrocategoria

- Quando invece non possiede questa proprietà applica l'algoritmo di Djikstra per trovare il percorso minimo, poi itera sui nodi di tale percorso e assegna ad ognuna di loro il valore di FROMx opportuno.
- Utilizza il valore di FROMx appena recuperato o assegnato alla categoria come valore della distanza della categoria dalla macrocategoria.

- Utilizza le distanze delle categorie dalle macrocategorie per determinare l'assegnamento dell'articolo.

Nel caso delle tre categorie composte viste prima, che sono state aggiunte per il lavoro e non erano presenti nel grafo originale, il valore *FROMx* trovato viene diminuito di uno perché non siano penalizzate nel confronto con le altre.

Anche con questa ottimizzazione, però, il calcolo risulta troppo pesante. Eseguendolo si stima che per completare il lavoro sarebbero necessari vari mesi di elaborazione. Si deve quindi procedere a calcolare le proprietà *FROMx* preventivamente utilizzando questo algoritmo esposto in pseudocodice, dove la funzione $collegati(x)$ restituisce il set di nodi collegati a x con un arco *SUBCATEGORYOF* entrante o uscente, $is\_defined(x.y)$ dice se il nodo x ha la proprietà y e l'operatore & effettua la concatenazione di stringhe. L'algoritmo va eseguito per ogni macrocategoria m

- nome=m.nome
- A.push(m)
- count=0
- while(A is not empty)
  - foreach x in A
    - x.FROM&nome=count
    - foreach c in collegati(x)
      - if NOT is_defined(c.FROM&nome) B.push(c)
  - count=count+1
  - A.clear()
  - A.putAll(B)

  - B.clear()

In questo modo si assegnano velocemente le proprietà *FROMx* a tutte le categorie. È poi possibile eseguire l'algoritmo per il calcolo delle distanze mostrato prima, che stavolta sarà molto più veloce poiché tutte le categorie connesse al grafo avranno la proprietà *FROMx* assegnata.

Le tre categorie che sono state create prima vengono gestite variando l'algoritmo perché ne aggiunga il contenuto al set A già alla seconda riga, in modo da diminuire di 1 le distanze e bilanciare gli effetti della modifica sulle distanze topologiche.

Infine per ogni pagina si segue la seguente procedura in pseudocodice

- create set $APPARTENENZE < Node >$
- create map $APPARTENENZE2\,\langle Node, number\rangle$
- foreach cat in pagina.getCategorieAppartenenza()
  - define distanzaMin=Double.MAX_VALUE
  - foreach $\langle distanza, macrocategoria\rangle$ in cat.getDistanzeDaMacrocategorie()
    - if distanza>distanzaMin
      - continue
    - if distanza==distanzaMin
      - APPARTENENZE.push(macrocategoria)
    - if distanza<distanzaMin
      - APPARTENENZE.clear()
      - APPARTENENZE.push(macrocategoria)
      - distanzaMin=distanza
  - APPARTENENZE2.push($\langle APPARTENENZE, 1/APPARTENENZE.size()\rangle$)
- Utilizza APPARTENENZE2 per calcolare l'assegnamento dell'articolo

Si noti che APPARTENENZE2 è una mappa, che oltre a contenere una lista di macrocategorie assegna ad ognuna un'etichetta pari al numero di componenti di APPARTENENZE. Questa etichetta viene usata per assegnare a ogni macrocategoria una quota dipendente non solo dal numero di categorie che le sono vicini ma anche al numero di categorie equidistanti da più categorie, informazione che altrimenti andrebbe persa. Il coefficiente 1/APPARTENENZE.size() indica proprio la quota derivante da quella categoria, che è ripartita appunto tra le macrocategorie equidistanti, se ce ne sono, altrimenti va tutta alla macrocategoria più vicina.

Il valore *Double*.MAX_*VALUE* indica, in Java, il massimo valore che può essere assegnato a una variabile double, che è molto più alto dei valori su cui lavorerà l'algoritmo. La variabile viene inizializzata a questo valore in modo che dal primo confronto tra la variabile e una distanza risulti sempre minore la distanza.

Le appartenenze sono calcolate in percentuali basandosi sul contenuto di APPARTENENZE2, in modo che ad ogni categoria di appartenenza sia data una quota uguale alle altre del 100%, dividendo ulteriormente questa quota nel caso una delle categorie sia equidistante da più macrocategorie contemporaneamente.

Quindi, ad esempio, se un articolo appartiene a 3 categorie vicine a *Geography and places* e una vicina a *History and events* il programma scriverà nell'output che la pagina appartiene al 75% a *Geography and places* e al 25% a *History and events*.

Quando una categoria è equidistante da più macrocategorie queste vengono annotate tutte, come descritto nell'algoritmo, ma insieme a un coefficiente inversamente proporzionale al numero di macrocategorie equidistanti dalla categoria, in modo da ripartire le quote di assegnamento di una categoria tra le macrocategorie che hanno da essa la stessa distanza.

Quindi se una pagina A appartiene a una categoria C1 e C2, dove C1 è vicino alla macrocategoria M1 e C2 è equidistante dalle macrocategorie M1 e M2 verrà assegnata per quando riguarda C1 un'appartenenza di valore 1 a M1 e per quanto riguarda C2 un'appartenenza di valore 0.5 a M1 e 0.5 a M2, così la pagina verrà assegnata a M1 per il 75% (ossia $\frac{0.5+1}{2}$) e a M2 per il 25% (ossia $\frac{0.5}{2}$).

Il risultato dell'elaborazione è un elenco di pagine nel formato

ID>Nome>M1:3 3 . 3 3 ;M2:6 6 . 6 6 ;M3:3 3 . 3 3 ;

Dove M1 e M2 sono le macrocategorie di appartenenza seguite dopo i due punti dai valori percentuali di importanza, troncati alla seconda cifra decimale.

4.4 Calcolo dell'affinità tra le macrocategorie

Un altro dato interessante che si può ottenere è la tendenza delle macrocategorie a sovrapporsi, cioè a contenere gli stessi articoli, che per ogni coppia di macrocategorie si può esprimere con un coefficiente di sovrapposizione.

Un metodo efficace per calcolare questi coefficienti è applicare la tecnica della *cosine similarity*.

Questa tecnica, molto usata nel data mining per stimare la vicinanza semantica di due testi, calcola il prodotto tra i vettori delle frequenze delle parole nei testi normalizzandolo con una divisione per il prodotto della norma dei vettori stessi.

$$cos(A, B) = \frac{A \bullet B}{||A|| * ||B||}$$

In questo caso i vettori con cui si rappresentano le macrocategorie sono composti da un elemento per ognuno degli n articoli, il cui valore è la percentuale di appartenenza dell'articolo alla macrocategoria, che vale 0 se l'articolo non appartiene affatto alla macrocategoria.

Di conseguenza il coefficiente di similarità tra due macrocategorie è la somma dei prodotti della percentuale di appartenenza a una e all'altra di ogni articolo divisa per la radice della somma dei quadrati delle percentuali di appartenenza:

$$cos(A, B) = \frac{\sum_{k=1}^{n} A(k) * B(k)}{\sqrt{\sum_{k=1}^{n} A(k)^2} * \sqrt{\sum_{k=1}^{n} B(k)^2}}$$

Questo valore è sempre positivo essendo tutte le percentuali maggiori di 0.

L'operazione di *cosine similarity*, naturalmente, fornisce lo stesso risultato invertendo l'ordine degli argomenti, e fornisce il valore massimo, ossia 1, quando si calcola la similitudine di un vettore rispetto a se stesso.

Il conteggio degli articoli assegnati a un certo numero di macrocategorie è semplice da realizzare, basta scorrere il file degli assegnamenti e contare il numero di articoli assegnati a esattamente n macrocategorie, per ogni n compreso tra 1 e il numero di macrocategorie totali.

4.5 Ricerca degli anelli di categorie

Per ricercare gli anelli all'interno del grafo si è usato l'algoritmo di Tarjan[1] per la ricerca delle componenti fortemente connesse, che quando sono composte da più di un nodo contengono dei cicli.

Chiamando V il numero dei nodi e E il numero degli archi, questo algoritmo riesce, in tempo $O(V + E)$, a elencare tutti gli insiemi distinti di nodi tali che ogni nodo sia raggiungibile da tutti gli altri percorrendo solo archi tra nodi dell'insieme.

L'algoritmo di Tarjan implementato ha questa struttura:

- create global stack S (visibilità globale)

- define global index=0 (visibilità globale)
- foreach n in Graph
  - if is_article(n) OR is_empty(n.getCategorieContenute())
    - continue;
  - if NOT is_defined(n.index)
    - tarjan(n)

La funzione tarjan(n) chiamata agisce per ricorsione ed ha questa struttura:

- n.index=global.index (assegna a n.index il valore della variabile globale index)
- n.lowlink=n.index
- global.index=global.index+1;
- S.push(n)
- foreach vp in n.getCategorieContenute()
  - if NOT is_defined(vp.index)
    - tarjan(vp)
    - n.lowlink=min(n.lowlink,vp.lowlink)
  - else
    - if S.contains(vp)
      - n.lowlink=min(n.lowlink,vp.index)
  - if n.index==n.lowlink
    - conta=0
    - while((estratto = S.pop()) != n)
      - print estratto
    - print ————————-" (serve all'utente a distinguere le diverse componenti)

Il risultato è un file contenente le liste dei nomi dei nodi categoria che formano degli anelli, separate da delle linee di soli trattini che fungono da separatori.

4.6 Normalizzazione dei baricentri

La normalizzazione dei baricentri avviene moltiplicando le proprietà *FROMx* assegnate a tutte le categorie con il metodo illustrato per degli opportuni coefficienti che portino i baricentri delle curve delle distribuzioni delle frequenze delle distanze dalle macrocategorie ad essere uguali tra di loro.

Bisogna innanzitutto calcolare i baricentri di queste curve. Chiamando D_m il vettore che ha come k-simo elemento il numero di categorie a distanza k dalla macrocategoria m, la formula per il baricentro è:

$$B_m = \frac{\sum_{x=1}^{t} D_m * x}{\sum_{x=1}^{t} D_m}$$

dove t è la massima distanza dalla macrocategoria in esame.

È possibile anche calcolare il baricentro dando ad ogni categoria un peso uguale al numero di articoli che contiene.

Se esistono N categorie, chiamando $n(c)$ il numero di articoli contenuti nella x-sima categoria e $d_m(c)$ la distanza della categoria da una macrocategoria m allora il baricentro di una macrocategoria calcolato in base al numero di articoli sarà:

$$B_m = \frac{\sum_{x=1}^{N} d_m(x) * n(x)}{\sum_{x=1}^{t} n(x)}$$

Per normalizzare i baricentri è sufficiente iterare sui nodi categoria e dividere le proprietà *FROMx* per i rispettivi baricentri, in modo che calcolando quelli nuovi si ottenga sempre 1.

Una volta modificate le proprietà *FROMx* è possibile rieffettuare l'assegnamento delle pagine alle macrocategorie e valutarne la correttezza, come si farà per tutte le altre varianti.

4.7 Percorso minimo seguendo l'orientamento degli archi

È possibile modificare l'algoritmo di assegnazione delle proprietà *FROMx* in modo che l'assegnamento proceda tenendo conto della direzione della relazione *BELONGSTO*, in questo modo:

- foreach m in macrocategorie
  - create set A=[m]
  - create set B

- define c=0
- LABEL ciclo
- foreach n in A
  - n.contatore=c
  - foreach v in n.getCategorieContenute()
    - if NOT is_defined(v.FROM&m.name) B.add(v)
- A.clear()
- A.addAll(B)
- B.clear()
- if NOT A.empty() GOTO ciclo

È possibile anche usare un algoritmo ricorsivo, basato su una funzione $f(n, d, m)$ dove n è un nodo in esame e d la distanza dalla macrocategoria di nome m in esame.

Tale funzione effettuerà questa semplice operazione:

- if n.FROM&m.name¡d AND is_defined(n.FROM&m.name)
  - return
- n.FROM&m.name=d
- foreach t in n.getCategorieContenute()
  - $f(t, d + 1, m)$

Si noti che è necessario gestire il caso che i nodi abbiano già la proprietà *FROMm*, perché il programma potrebbe giungere a uno di essi prima tramite un percorso lungo e in seguito con uno più corto. L'algoritmo è quindi più semplice ma meno efficiente.

Una volta assegnate le proprietà *FROMx* si può procedere all'assegnamento delle pagine alle macrocategorie con il metodo di prima, con la differenza che non tutte le categorie avranno tutte le proprietà *FROMx*, anzi, generalmente ne avranno due o tre. Si deve quindi modificare l'algoritmo per ignorare le proprietà *FROMx* mancanti, come se le macrocategorie fossero a distanza infinita dai nodi categoria privi delle rispettive proprietà *FROMx*.

4.8 Spostamento dei baricentri con percorsi diretti

Una volta assegnate le proprietà FROMx come illustrato si calcola il baricentro delle curve di distribuzione di frequenza di tali proprietà e si sottrae a ogni valore di queste proprietà il valore del baricentro della curva riguardante la macrocategoria da cui rappresenta la distanza topologica, quindi si ricalcolano gli assegnamenti.

Le ragioni dell'uso della sottrazione al posto della moltiplicazione verranno illustrate nei risultati.

4.9 Costo di attraversamento differenziato in base all'orientamento

Un'altra possibilità è quella di dare alle relazioni un costo di attraversamento diverso a seconda del modo in cui sono attraversate. Si potrebbe ad esempio dare un costo di 3 al passaggio in senso contrario alla relazione e di 1 al passaggio lungo l'orientamento di *SUBCATEGORYOF*.

Per assegnare le proprietà *FROMx* in questo modo è necessario utilizzare delle mappe al posto dei set, in modo che ad ogni nodo esplorato si possa assegnare una distanza specifica. L'algoritmo è il seguente:

- Create A$< Node, distanza >$
- Create B$< Node, distanza >$
- foreach m in macrocategorie
  - A.clear()
  - A.put($< m, 0 >$)
  - LABELciclo
  - foreach $< n, d >$ in A
    - foreach t in n.getCategorieContenute()
      - if NOT t.FROM&m.name < n.FROM&m.name B.put($< t, d + 1 >$)
    - foreach t in n.getCategorieAppartenenza()
      - if NOT t.FROM&m.name+2 < n.FROM&m.name B.put($< t, d + 3 >$)
  - foreach $< n, d >$ in B

        * n.FROM&m.name=d
    - A.clear()
    - A.putAll(B)
    - B.clear()
    - if A.size()>0 GOTO ciclo

Rispetto all'algoritmo di prima si utilizzano delle mappe per assegnare dei costi specifici e, soprattutto, si gestisce la possibilità che il nodo nuovo abbia già la proprietà *FROMx* ma con valore maggiore di quello che si vorrebbe assegnare.

Questa modifica è necessaria poiché, differenziando i costi, si potrebbe raggiungere subito un nodo e assegnarli una distanza incrementata di 3 e, dopo uno o due passi, raggiungere lo stesso nodo con passaggi orientati a costo 1 e evitare di assegnare la proprietà *FROMx* che gli spetterebbe.

4.10 Assegnamento maggioritario

L'assegnamento di ogni pagina a una sola macrocategoria si può effettuare calcolando le quote di appartenenza come nel caso base e assegnando la pagina alla macrocategoria che ha la quota di appartenenza maggiore.

Per farlo è sufficiente iterare sul file degli assegnamenti e determinare per ogni riga la quota maggiore, quindi scrivere il risultato in un nuovo file. Nel caso ci siano due o più categorie che possiedono quote uguali e maggiori delle altre vengono scritte insieme al nome dell'articolo in un altro file.

4.11 Assegnamento con ripartizione di punteggi

Esistono diversi algoritmi per calcolare il flusso tra due nodi. Per esempio è possibile considerare ogni arco come una resistenza, ignorandone l'orientamento, calcolare il flusso tra due nodi come se fosse l'inverso della resistenza ottenuta applicando le formule per le resistenze in parallelo o in serie. Il problema è che praticamente impossibile applicare questi algoritmi in un grafo così grosso, perché richiederebbero un tempo di elaborazione eccessivo. Si è utilizzata quindi un'euristica basata sulla ripartizione di un punteggio tra le categorie contenute in una macrocategoria, che agisce ricorsivamente.

Innanzitutto è necessario eliminare i cicli, che impedirebbero all'algoritmo di terminare.

Per farlo è sufficiente iterare sull'elenco delle strutture fortemente connesse generato precedentemente grazie all'algoritmo di Tarjan e, per ogni

nodo di ognuna di esse eliminare le relazioni che portano ad altri nodi della stessa struttura e che quindi portano a generare dei cicli.

Una volta eliminati, si applica questo algoritmo:

- foreach m in macrocategorie
  - assegna(m, 10000)

Il valore 10000 è assolutamente arbitrario, qualsiasi costante positiva uguale per tutte le macrocategorie è sufficiente.

La funzione $assegna(n, d)$ è così definita

- if is_defined(n.DIST&m.name)
  - n.DIST&m.name =n.DIST&m.name+d
- foreach c in n.getCategorieContenute()
  - assegna(c, $\frac{d}{n.getCategorieContenute().size()}$)

Quindi ad ogni categoria raggiungibile a partire da una macrocategoria m seguendo gli archi nel loro orientamento viene assegnata la proprietà DISTm. Non si è usato il nome FROMx perché questo valore ha un diverso significato e utilizzo.

Infatti si notano le seguenti differenze:

- Se una categoria ne contiene solo un'altra, entrambe avranno lo stesso valore DISTm, che non è influenzato da catene di nodi collegati in sequenza con un solo arco entrante e uno uscente.
- Sono penalizzate le categorie che sono contenute insieme a molte altre in una categoria più grande
- Sono premiate le categorie che sono raggiungibili con più percorsi diretti e distinti, poiché i valori DISTm si sommano

4.12 Assegnamento con probabilità di raggiungere la macrocategoria

È possibile calcolare per ogni nodo le probabilità di raggiungere una certa macrocategoria muovendosi solo in direzione dell'orientamento degli archi e scegliendone uno a caso qualora ce ne fossero più di uno entranti.

Posto che la probabilità che ha una macrocategoria di essere raggiunta da se stessa sia 1, si definisce la probabilità di raggiungere la macrocategoria da un nodo n come la somma delle probabilità dei nodi di partenza degli archi uscenti da n diviso il numero di archi uscenti da n totali. Se per esempio un nodo avesse tre archi entranti (cioè rappresentasse una categoria contenuta in altre tre categorie) e due di essi avessero una probabilità di essere raggiunti di 0.5 e 0.3, mentre il terzo non ha tale proprietà e quindi viene considerata nulla, le probabilità di raggiungiere la macrocategoria partendo dalla categoria sarebbero:
$\frac{0.5+0.3+0}{3} = 0.267$

La probabilità di raggiungimento di una categoria è dunque calcolata in base a quelle delle categorie che la contengono, quindi è immediato pensare a un algoritmo ricorsivo. Tuttavia, poiché un nodo può essere analizzato più volte (possono esserci dei bivi e dei ricongiungimenti nei percorsi diretti), ogni volta si dovrebbe ricalcolare il suo valore di probabilità e quello di tutti i nodi raggiungibili da esso. Sarebbe conveniente, quindi, procedere con la ricorsione partendo da un nodo n solo quando tutti i valori di probabilità delle categorie che contengono n sono stati calcolati, in modo da avere la sicurezza che non sarà più necessario analizzare nuovamente n. Tuttavia, poiché si procede seguendo la direzione degli archi, non è possibile sapere immediatamente se le categorie che contengono n saranno poi analizzate o invece non saranno mai raggiunte. Si deve quindi determinare prima questa informazione utilizzando l'assegnamento delle proprietà FROMx secondo l'orientamento degli archi visto prima. Si seguono queste tre fasi:

- Individua e elimina i cicli dal grafo
- Assegna le proprietà FROMx alle categorie secondo l'orientamento degli archi. Serviranno per individuare i nodi raggiungibili dalle macrocategorie con percorsi diretti, i valori FROMx effettivamente assegnati non vengono però usati.
- Itera sulle macrocategorie, e per ogni macrocategoria t chiama $elabora(t, t.name)$

La funzione $elabora(n, name)$ è così definita:

- define totale=0, entranti=0, incompleto=false
- entranti=n.getCategorieAppartenenza().size()
- foreach c in n.getCategorieAppartenenza()
  - if is_defined(c.FROM&name) AND is_defined(c.REACH&name)

        * totale=totale+c.REACH&name
    - if is_defined(c.FROM&name) AND NOT is_defined(c.REACH&name)
        * incompleto=true

- if totale==0
    - totale=1, entranti=1
    - incompleto=false
- if NOT incompleto
    - n.REACH&name=$\frac{totale}{entranti}$
    - foreach h in n.getCategorieContenute()
        * elabora(h,name)
- if incompleto
    - return

Si tratta quindidi un algoritmo depth-first modificato per non analizzare i nodi che analizzerebbe comunque una seconda volta, comportandosi in maniera simile a una ricerca breadth-first.

Una volta assegnati i valori *REACHx* si itererà sui nodi pagina assegnandoli alle macrocategorie in base ai valori maggiori delle proprietà *REACHx* delle categorie che li contengono, come fatto prima con *DISTx*.

Capitolo 5

Valutazione dei risultati

5.1 Analisi statistica del grafo

Una volta generato il grafo lo si esporta nel formato usato da Pajek, per poterlo poi analizzare con igraph ed estrarne delle proprietà statistiche globali.

Una di queste è il diametro, ossia la massima distanza topologica che si può trovare tra due nodi del grafo, che viene calcolato con il comando diameter. Si consideri che la distanza topologica è considerata seguendo gli archi solo in base al loro orientamento. Si determina così che il diametro del grafo è di 32 nodi, e che i nodi con il più lungo percorso minimo sono sono *Prehistoric life sorted by geography* (Una categoria vuota sugli animali preistorici organizzati in base al luogo) e *BMW M20* (una voce su un'automobile). Poiché il nodo rappresentante una pagina può essere collegato solo a nodi rappresentanti categorie si calcola immediatamente che il più lungo percorso minimo tra categorie è lungo 31 nodi. Inoltre, la distanza media tra due nodi qualsiasi è 5.5568 e la densità del grafo, ossia il rapporto tra gli archi esistenti e tutti gli archi possibili, è $6.28 * 10^{-7}$.

La densità di un grafo è definita come:

$$\frac{a}{(p+c)^2}$$

dove a è il numero di archi, inizialmente circa 40 milioni mentre p è il numero di pagine e c il numero di categorie, dunque $p + c$ valeva, prima del filtraggio, circa 7.5 milioni. La densità prima del filtraggio era $8.16 * 10^{-7}$, e si è abbassata rimuovendo i nodi e gli archi delle categorie non semantiche. Questo cambiamento indica che i nodi privi di significato semantico hanno un livello di interconnessione maggiore rispetto a quelli semantici, quindi rimuovendoli si ottiene un grafo meno denso.

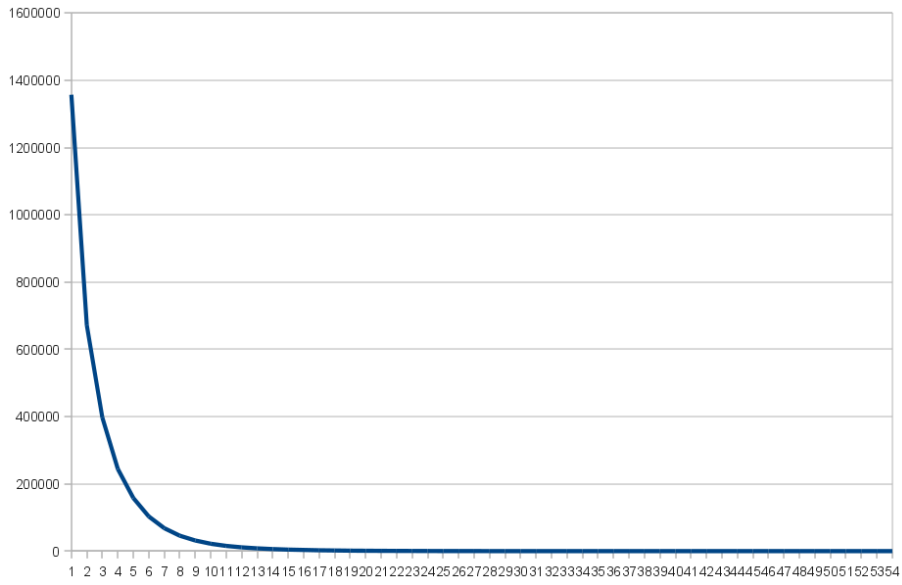

Figura 5.1: Il numero di pagine assegnate a una certa quantità di categorie. È evidente come la maggior parte delle pagine sia assegnata a poche categorie, infatti il 64% degli articoli appartiene a una o due categorie.

È possibile estrarre delle informazioni anche sul numero di pagine assegnate a una certa quantità di categorie utilizzando dei programmi creati appositamente. Si determinano così i dati indicati in tabella, riguardanti le pagine assegnate a meno di 7 categorie, che costituiscono il 93% del totale:

1	1356845
2	671618
3	397437
4	244293
5	157926
6	103100

I dati possono essere illustrati graficamente (Figura 5.1).La media delle categorie a cui è assegnata ogni pagina è 2.68. La pagina assegnata a più categorie, ben 70, è *Winston Churchill*, seguita da *English language*, contenuto in 60 categorie, *Imperata cylindrica*, pianta la cui pagina è assegnata a 59 categorie, e poi *Panicum virgatum* (una pianta diffusa nelle praterie degli Stati Uniti) che è assegnata a 56 categorie come l'articolo *Albert Einstein*.

5.2 Assegnamento con il metodo di Kittur

L'operazione di assegnamento delle proprietà FROMx ha richiesto circa 17 ore, mentre quella di assegnamento delle pagine alle macrocategorie è durata circa 35 ore. Una volta terminata l'esecuzione del programma si ha un file contenente gli assegnamenti degli articoli alle macrocategorie.

Di seguito si riportano alcuni esempi di assegnamenti contenuti nel file:

```
Milan:History and events:100;
Italy:Culture:30;History and events:70;
Politecnico di Milano:Education:87,5;History and
    events:12,5;
Java (software platform):Computing:50;Technology and
    applied sciences:50;
Java:Geography and places:100; (L'isola principale
    dell'Indonesia)
Earth:Geography and places:100;
Albert Einstein>Society:2,93;Politics:11,39;History
    and events:5,31;Technology and applied sciences
    :18,53;Science:18,53;People:24,78;Mathematics
    :18,53;
Nintendo:Agriculture:6,67;Culture:6,67;History and
    events:71,67;Geography and places:6,67;Philosophy
    :6,67;Technology and applied sciences:1,67; (famosa
     azienda produttrice di videogiochi)
20th Century Boys: Environment:13,39;Society:13,39;
    Culture:13,39;Education:13,39;History and events
    :6,25;Sports:13,39;Technology and applied sciences
    :13,39;Science:13,39; (fumetto giapponese di
    fantascienza)
Semiconductor fuse:Business:50;Science:50;
```

L'assegnamento di *Nintendo* in quote uguali a *Agriculture* e *Technology and applied sciences* è un esempio di anomalia che si vorrebbe evitare e avviene a causa dei percorsi illustrati nelle Figure 5.2 e 5.3. Questi errori sono da analizzare per poter migliorare l'algoritmo e renderlo più preciso.

È possibile effettuare una statistica delle dimensioni di ogni macrocategoria sommando i valori percentuali degli assegnamenti effettuati a ogni macrocategoria.

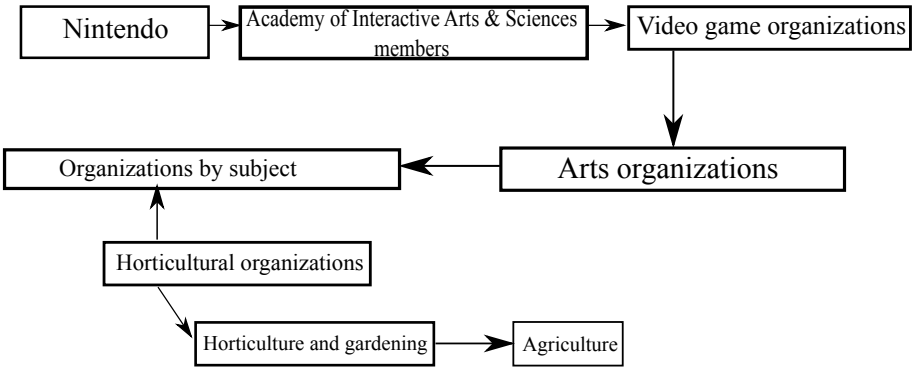

Figura 5.2:
Il percorso minimo che collega la pagina Nintendo alla macrocategoria *Agriculture*. Le frecce indicano la direzione della relazione di assegnamento, quindi *Arts organizations* è contenuta in O*rganizations by subject* e contiene *Video game organizations*. Si osserva che il passaggio a *Organizations by subject* potrebbe portare indirettamente a molte categorie su argomenti totalmente diversi.

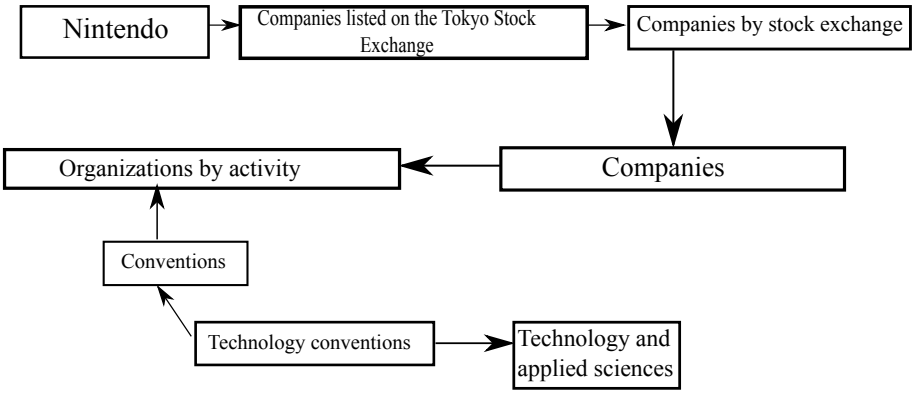

Figura 5.3:
Il percorso minimo tra la pagina *Nintendo* e la categoria *Technology and applied sciences*, che ha la stessa quantità di passaggi di quello tra *Nintendo* e *Agriculture.*

Geography and places	60337175,28
History and events	57327498,44
Culture	32730545,69
People	28247194,2
Agriculture	19682932,19
Sports	15642414,42
Society	10686177,42
Politics	8703228
Technology and applied sciences	8290477,04
Education	7904696,77
Law	5856760,26
Environment	4541602,99
Business	3455513,21
Science	3424014,85
Language	3186144,1
Mathematics	2789931,76
Belief	2504420,2
Health	2164970,09
Philosophy	1846347,75
Computing	1662195,38
Arts	1231224,49

La macrocategoria più grossa è *Geography and places*, come era prevedibile vista la frequenza con cui si ottengono pagine su luoghi geografici utilizzando la funzione una pagina a caso.

Questo accade soprattutto perché esistono dei bot che generano rapidamente migliaia di pagine su paesi e comuni molto piccoli a partire da database pubblici e basandosi su un template.

Vengono subito dopo pagine di argomento storico, culturale e infine biografico. *Agriculture* è molto diffusa perché le sue sotto-categorie sono molte e ne contengono a loro volta altre in una quantità più grande di quanto avvenga in genere con il resto del grafo, quindi esistono molte categorie a bassa distanza da *Agriculture* che causano degli assegnamenti anomali. Questo fatto verrà approfondito nel prossimo capitolo.

È da osservare che nonostante esistano numerosissime biografie, come si può notare scorrendo il file degli assegnamenti, la macrocategoria *People* non è subito dopo *Geography*. La causa è probabilmente il fatto che ogni personalità enciclopedica è famosa per qualcosa, per esempio un attore rientrerà nella macrocategoria *Arts* o uno sportivo in *Sports*, rendendo molto raro che la macrocategoria *People* riceva assegnamenti con percentuali alte.

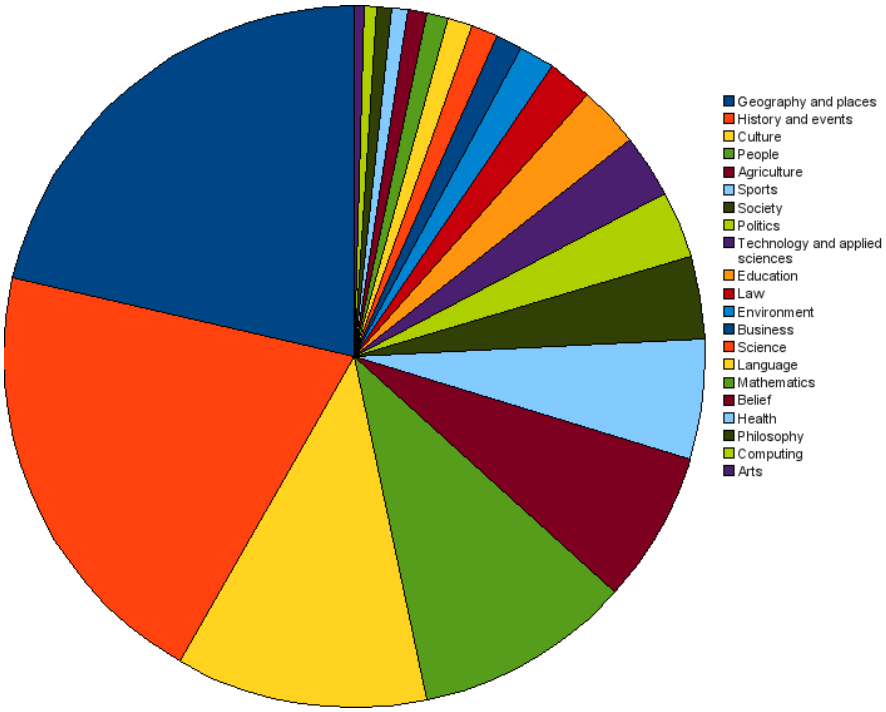

Figura 5.4:
Diagramma a torta che rappresenta l'estensione delle macrocategorie classificate con questo metodo

Se a ciò si unisce la struttura molto tassonomica della macrocategoria, che è organizzata in maniera gerarchica distribuendo le persone in base al luogo e all'anno di nascita passando per vari livelli di categorie sempre più specifiche il fatto che *People* sia quarta in classifica è più comprensibile.

Si possono rappresentare i risultati anche con un diagramma a torta (Immagine 5.4)

La sovrapposizione tra le macrocategorie, ossia la tendenza di due macrocategorie a ricevere l'assegnazione delle stesse pagine, può ora essere calcolata con il metodo della cosine similarity.

Possiamo vedere, per ogni macrocategoria, qual'è quella con cui ha il maggiore coefficiente di similarità dopo se stessa:

Science	Mathematics	0,35
Mathematics	Science	0,35
Technology_and_applied_sciences	Science	0,21
Society	Health	0,20
Health	Society	0,20
Culture	People	0,17
People	Culture	0,17
History_and_events	Culture	0,14
Agriculture	Culture	0,12
Geography_and_places	Culture	0,11
Business	Technology_and_applied_sciences	0,10
Politics	People	0,10
Belief	Culture	0,09
Sports	People	0,08
Computing	Technology_and_applied_sciences	0,07
Philosophy	Mathematics	0,07
Law	History_and_events	0,07
Education	Science	0,06
Language	Geography_and_places	0,06
Arts	Language	0,06
Environment	Sports	0,05

Il coefficiente più alto è quello tra *Mathematics* and *Science*, seguito dalla coppia *Science* e *Technology* and applied sciences.

La vicinanza semantica tra *Mathematics* e *Science* era risultata anche nella mappa visuale ottenuta da Holloway[20] usando un metodo diverso.

Si osserva inoltre che *Sport* ha come categoria più simile *People*, grazie alle numerose pagine sugli atleti che rientrano in entrambe le categorie.

Questo risultato era prevedibile, vista la vicinanza semantica tra i due argomenti, mentre non è molto chiaro perché la macrocategoria più simile a *Environment* sia *Sports* e non, come verrebbe da pensare, *Geography* oppure *Science*.

Scorrendo allora il file degli assegnamenti per capire questa anomalia si notano numerose pagine assegnate al 50% a *Sports* e al 50% a *Environment*, per esempio l'articolo *Coluber hortulanus*.

Le catene di collegamenti verso le due macrocategorie sono illustrate nell'immagine 5.6.

L'articolo parla di un serpente e viene collegato con vari passaggi intermedi alla categoria *Vertebrates* e da qui a *Birds*. *Birds* è vicina a *Sports* per l'hobby di dare da mangiare agli uccelli e vicina a *Environment* per la

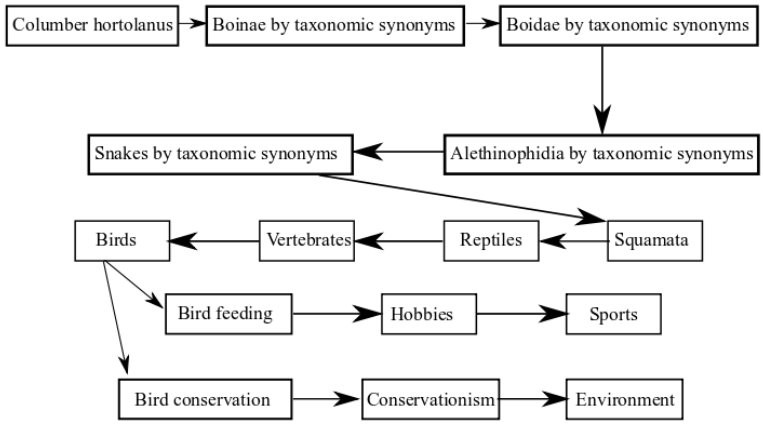

Figura 5.5:
Il percorso minimo tra *Columber hortolanus* e le macrocategorie *Sports* e *Environment*

protezione delle specie volatili.

Quindi qualsiasi pagina su un rettile che non sia abbinata a particolari categorie ma rientri solo nella normale categorizzazione tassonomica dei rettili viene assegnata automaticamente a *Sports* e *Environments*, generando questo risultato inaspettato.

Un altro risultato curioso è la similitudine tra *Society* e *Health*. Come prima, basta guardare il file degli assegnamenti per trovare molti articoli assegnati a entrambe al 50%.

Un esempio è l'articolo *Sea lemon*, che parla di un mollusco noto in italiano come *Doridídae*, che ha i collegamenti illustrati nella Figura 5.6

Proprio come *Sports* e *Environment* si ha la categoria sui molluschi che è vicina a entrambe queste macrocategorie, ed essendo la classificazione delle pagine di biologia generalmente strutturata in modo gerarchico (ricalcando la tassonomia di Linneo) si hanno centinaia o migliaia di pagine su specie di molluschi assegnate a *Society* e *Health* in uguale misura, generando una vicinanza delle due macrocategorie.

Si può ora passare ad osservare gli indicatori di similitudine più bassi per ogni macrocategoria

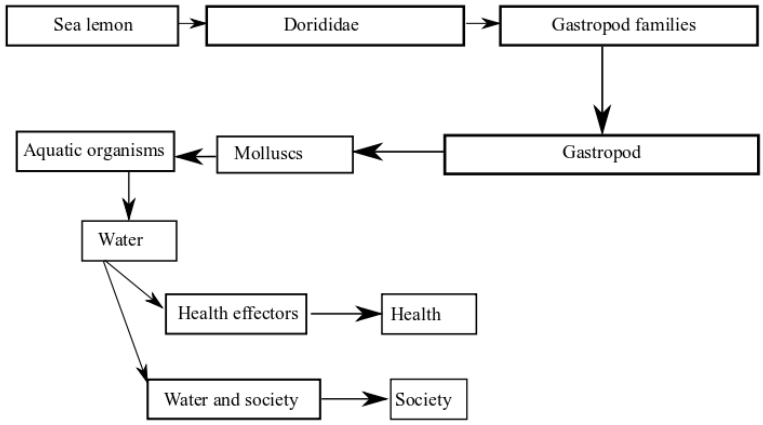

Figura 5.6:
Il percorso minimo tra *Sea lemon* e le macrocategorie *Health e Society*

Technology_and_applied_sciences	Belief	0,01755
Philosophy	People	0,00988
Society	Arts	0,00935
Geography_and_places	Health	0,00909
Language	Politics	0,00830
Sports	Health	0,00746
Business	Belief	0,00701
Agriculture	Arts	0,00675
History_and_events	Health	0,00607
Culture	Computing	0,00537
Science	Arts	0,00525
Belief	Health	0,00363
Environment	Politics	0,00309
Mathematics	Health	0,00281
Education	Computing	0,00226
Politics	Computing	0,00207
People	Health	0,00159
Law	Arts	0,00145
Arts	Computing	0,00107

Le due macrocategorie più distanti sono *Computing* e *Arts* seguite dalla coppia *Law* e *Arts*. È possibile individuare la categoria più multidisciplinare

con un' euristica che indichi, per ognuna, la somma di tutti i valori di cosine similarity con le altre. I valori ottenuti in questo modo sono:

Science	1.25
Culture	1.24
Technology and applied sciences	1.04
History and events	1.00
Society	0.96
Mathematics	0.92
People	0.90
Geography and places	0.74
Agriculture	0.71
Business	0.64
Politics	0.59
Philosophy	0.58
Sports	0.57
Law	0.55
Education	0.53
Environment	0.53
Language	0.53
Health	0.50
Belief	0.43
Computing	0.30
Arts	0.30

La macrocategoria più interdisciplinare, secondo questo criterio empirico, è *Science*, seguita da *Culture*. Quelle meno legate alle altre sono invece *Computing* e *Arts*.

5.3 Valutazione della precisione degli assegnamenti

Per valutare la precisione degli assegnamenti ottenuti con questo metodo si è scelto di effettuare l'assegnamento da una persona e confrontare i risultati con quelli forniti dal programma. Kittur e Suh[15] hanno fatto questo esperimento facendo ripartire 100 punti rappresentanti la correlazione tra le 11 macrocategorie per vari articoli a un campione di operatori reclutati su internet.

Questo metodo però richiede molto tempo, sia per preparare il quiz che per aspettare che qualcuno accetti il lavoro e lo compia, che per controllare che i risultati non siano stati inseriti a caso.

Quindi si userà un metodo più semplice: un programma estrae delle righe a caso dal file degli assegnamenti e chiede all'utente di assegnare la pagina estratta, senza mostrare la valutazione effettuata dal programma per non influenzarlo. Quindi viene calcolato un indice di correttezza sulla base del confronto tra i punteggi decisi dall'utente e l'assegnazione fatta automaticamente dal programma.

Per valutare la similitudine tra l'assegnamento svolto dalla persona e quello svolto dall'algoritmo si utilizza nuovamente la cosine similarity.

Dunque chiamando A_u il vettore degli assegnamenti alle macrocategorie svolto dal valutatore umano e A_a l'analogo vettore prodotto dall'algoritmo e assumendo che il k-simo elemento di ogni vettore corrisponda alla stessa macrocategoria e abbia valore pari alla percentuale dell'assegnamento, o 0 se non è stato assegnato alla categoria numerata con k, si ha che il valore che indica la correttezza dell'assegnamento è dato da:

$$cos(A_a, A_u) = \frac{\sum_{k=1}^{n} A_u(k) * A_a(k)}{\sqrt{\sum_{k=1}^{n} A_a(k)^2} * \sqrt{\sum_{k=1}^{n} A_u(k)^2}}$$

che è semplicemente la formula della cosine similarity applicata al nuovo caso.

Il risultato della valutazione, effettuata con 50 articoli, è 0.34.

Questo dato non è confrontabile con quello dei due studiosi per vari motivi:

- Il numero di macrocategorie scelte è quasi raddoppiato, sono passate da 11 a 21.
- Anche il numero delle pagine e delle categorie esistenti è cresciuto notevolmente. Le categorie sono più che raddoppiate e le pagine sono passate da 2 a 3 milioni.
- L'assegnamento manuale è stato effettuato da una sola persona e non da più volontari distinti.
- Le pagine sono state selezionate casualmente fra tutti gli articoli e non solo tra gli articoli in vetrina.

È importante sottolineare che il valutatore ha la possibilità di saltare delle valutazioni se non riesce a stabilire l'argomento della pagina, quindi le 50 pagine valutate sono le rimanenti di un set più grande da cui sono state tolte delle pagine scartate dall'utente.

Poiché le pagine sono state selezionate a caso, esiste la possibilità che questo risultato sia troppo aleatorio e non renda possibile ottenere una valutazione utile a comparare la precisione del criterio di assegnamento con

quella degli altri criteri, e quindi si vuole sapere se 50 assegnamenti sono sufficienti a valutare la correttezza degli assegnamenti senza risentire troppo delle differenze tra i set dei 50 articoli scelti. Dunque vengono estratte altre 50 pagine a caso e viene effettuato di nuovo il test. Il risultato è nuovamente 0.34, quindi è ragionevole considerare il risultato attendibile. Le macrocategorie identificate con più precisione sono state History and events, *Geography and places*, *Sports* e *People*. Ci sono stati 6 assegnamenti perfettamente uguali e 16 assegnamenti simili, ossia con un coseno compreso tra 0 e 1, mentre i rimanenti 38 sono stati totalmente diversi e hanno quindi ricevuto un valore di correttezza pari a 0. Il problema più grosso è la tendenza dell'algoritmo ad assegnare le pagine ad *Agriculture*, perché questa anomalia non è causata da categorie non semantiche o ambigue che interferiscono con il calcolo del percorso minimo, problema risolvibile semplicemente eliminandole o modificandole manualmente, ma da una bassa profondità del ramo delle categorie che è una caratteristica intrinseca dei dati di partenza.

L'analisi della frequenza degli assegnamenti a un certo numero di macrocategorie fornisce i seguenti risultati:

1	1024121
2	594362
3	406690
4	310713
5	164926
6	92970
7	54924
8	57375
9	52165
10	42071
11	13478
12	5347
13	1518
14	664
15	304
16	137
17	3
18	374
19	11
20	1

Come era prevedibile, all'aumentare del numero delle macrocategorie diminuiscono gli articoli assegnati, infatti la maggior parte degli articoli,

precisamente il 57% del totale, viene assegnato a una o due macrocategorie.

Esiste solo un articolo assegnato a 20 macrocategorie, e si tratta di *Paris ticket "t"*, una pagina riguardante un tipo di biglietto per i mezzi pubblici di Parigi, il cui assegnamento è

```
Paris ticket "t">Society:5,56;Environment:5,56;
    Agriculture:5,56;Culture:5,56;Business:5,56;History
     and events:5,56;Belief:5,56;Health:5,56;Sports
    :5,56;Science:5,56;Mathematics:5,56;Education:5,56;
    Geography and places:5,56;Language:5,56;Law:5,56;
    Philosophy:5,56;Computing:5,56;Technology and
    applied sciences:5,56;
```

L'assegnamento non è certamente corretto, visto che difficilmente si collegherebbe un biglietto per i mezzi pubblici di Parigi all'agricoltura, alla storia, alle religioni o allo sport.

È naturalmente possibile rappresentare graficamente questi risultati (Figura 5.7)

Si nota che la curva non è decrescente in prossimità del valore 8 e del valore 18.

È difficile risalire al motivo di questo comportamento vista la grande complessità del grafo, ma probabilmente è dovuto all'esistenza di una o più categorie che sono equidistanti da otto macrocategorie e causano l'assegnamento contemporaneo a tutte queste degli articoli contenuti al loro interno e nelle loro sotto-categorie (che però possono appartenere a loro volta ad altre categorie che cambiano i risultati).

Questa ipotesi si può verificare osservando che le pagine con 18 assegnamenti contemporanei sono quasi tutte relative ad argomenti economici, come *Currency*, *Redlining* (la pratica di alzare notevolmente i costi dei servizi in una certa area per discriminarne gli abitanti) o *Dollarization*.

Nel caso delle pagine con 8 assegnamenti le categorie causa dell'anomalia sono difficili da individuare perché coprono gli argomenti più disparati.

5.4 Ricerca dei cicli

La ricerca delle strutture fortemente connesse composte da più di un nodo ne rileva 93; ognuna di esse contiene almeno un ciclo distinto. Per la maggior parte sono coppie di categorie che si contengono reciprocamente, ma si trovano anche casi più complessi, ad esempio le categorie

```
Sources
```

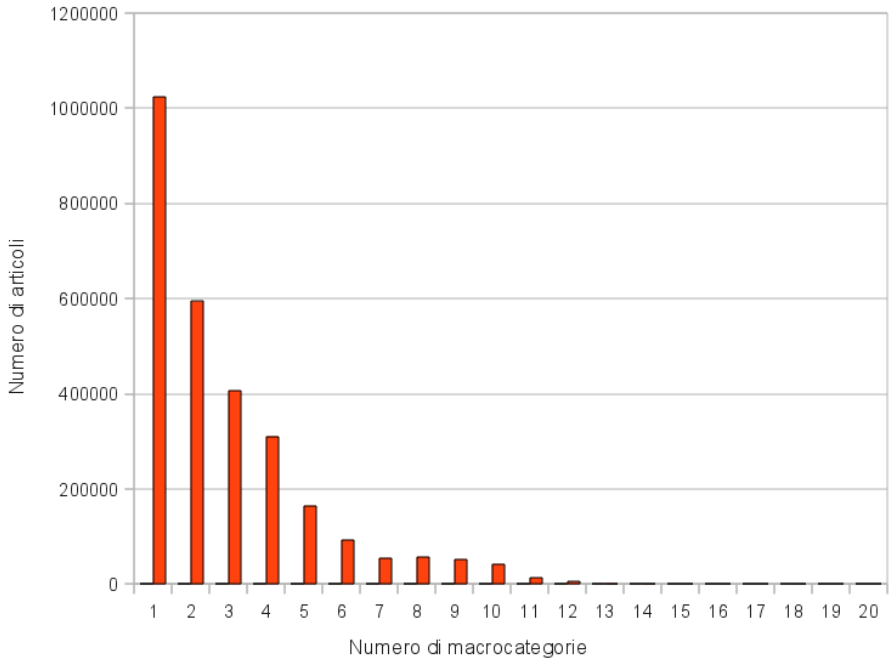

Figura 5.7: Il numero di articoli assegnati a un certo numero di macrocategorie. Si osserva che la curva è decrescente, ad indicare che la maggior parte degli articoli è assegnata a un piccolo numero di macrocategorie. L'unica eccezione è causata dalle pagine assegnate a 8 macrocategorie che sono in quantità leggermente maggiore di quelle assegnate a 7. Altra eccezione sono gli articoli a distanza 18 che sono molti di più di quelli a distanza 17.

Reference
Perception
Sources of knowledge
Knowledge
Memory

Formano una struttura fortemente connessa piuttosto complessa (Figura 5.8).

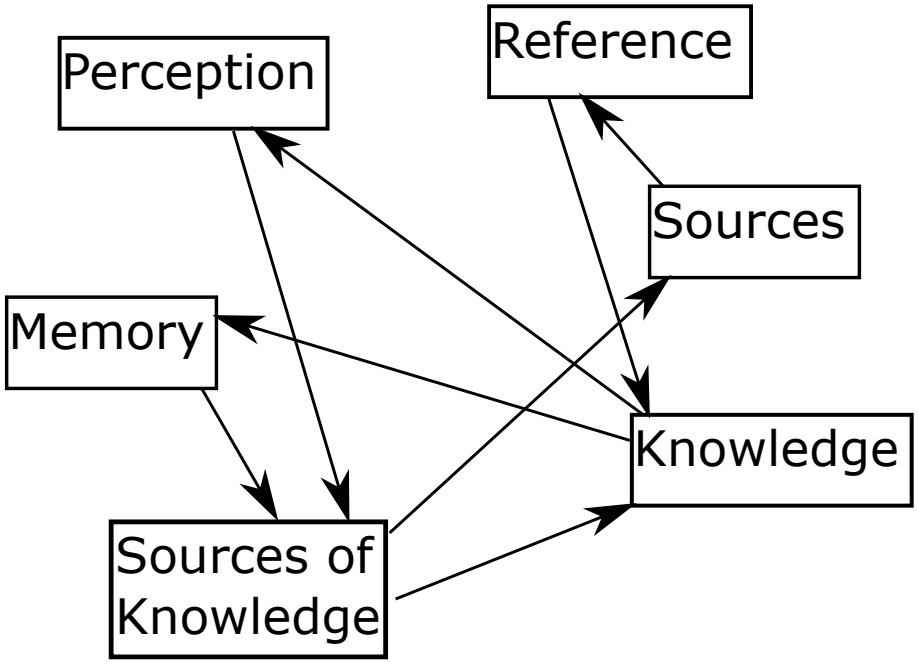

Figura 5.8: Una struttura fortemente connessa trovata all'interno del grafo. Partendo da ognuna delle sei categorie si possono raggiungere tutte le altre con un percorso che rimanga dentro la struttura. In questo caso si contano tre anelli distinti

Gli altri casi di anelli sono in massima parte dovuti a categorie sovrapposte, che indicano con un nome diverso concetti praticamente uguali, come

Microsoft server software
Microsoft server technology

oppure

Roads in Bangladesh
Road transport in Bangladesh

altri casi sono dovuti a degli errori degli utenti, come l'anello

```
Italian queens consort
Lombardic queens consort
```

dovuto all'errato inserimento della categoria *Italian queens consort* come sotto-categoria di *Lombardic queens consort.*

5.5 Normalizzazione dei baricentri

Per calcolare i baricentri si deve creare un programma che iteri sui nodi categoria e conti semplicemente la frequenza dei valori delle proprietà *FROMx*, per poi applicare la formula illustrata.

I baricentri sono

Computing	10.66
Health	10.22
Arts	10.19
Mathematics	10.01
Business	9.79
Language	9.73
Environment	9.68
Politics	9.55
Science	9.55
Law	9.53
Education	9.52
Belief	9.47
Technology and applied sciences	9.44
Sports	9.33
Agriculture	9.28
Philosophy	9.28
Society	9.02
People	8.82
History and events	8.55
Culture	8.52
Geography and places	8.38

Un valore alto indica che la categoria ha un'alta profondità tassonomica, mentre un valore basso indica che le categorie tendono a contenere molte altre categorie. *Agriculture*, come previsto, ha un baricentro basso, il quinto più basso nella lista.

Dando invece a ogni categoria un peso corrispondente alla quantità di articoli contenuti, i risultati sono

Computing	10.48
Health	10.12
Arts	10.04
Mathematics	9.87
Language	9.66
Environment	9.64
Business	9.59
Law	9.52
Education	9.43
Science	9.42
Belief	9.4
Politics	9.33
Technology and applied sciences	9.27
Sports	9.16
Philosophy	9.11
Agriculture	9.03
Society	8.86
People	8.57
Geography and places	8.49
History and events	8.42
Culture	8.34

La discrepanza dei valori dei baricentri ottenuti con i due metodi è molto bassa: quasi tutti i baricentri aumentano dell'1-2% se non si considera il numero di articoli contenuti in una categoria, a parte *Geography and places* che è l'unica a diminuire, di solo l'1%.

Confrontando le frequenze ottenute nei due modi è possibile anche calcolare il numero medio di articoli per categoria al variare della distanza dalle macrocategorie, che è semplicemente il rapporto tra il numero di categorie e il numero di articoli che si trovano a una certa distanza da una certa macrocategoria.

Si ottiene così un grafico (Figura 5.9) dal quale appare evidente perché *Agriculture* abbia ricevuto così tanti assegnamenti: alla distanza di 3 passi ha una media di 54 assegnamenti per categoria, la più alta fra tutte, così come a 8 passi di distanza riceve una media di 23.27 assegnamenti per categoria, ancora una volta una delle più alte.

Osservando il grafico della distribuzione di frequenza delle distanze delle categorie dalle macrocategorie (Figura 5.10), usando il numero degli articoli contenuti in ognuna come peso, si può avere un'indicazione ancora più evidente del fatto che *Agriculture* riceve molti assegnamenti per la sua bassa

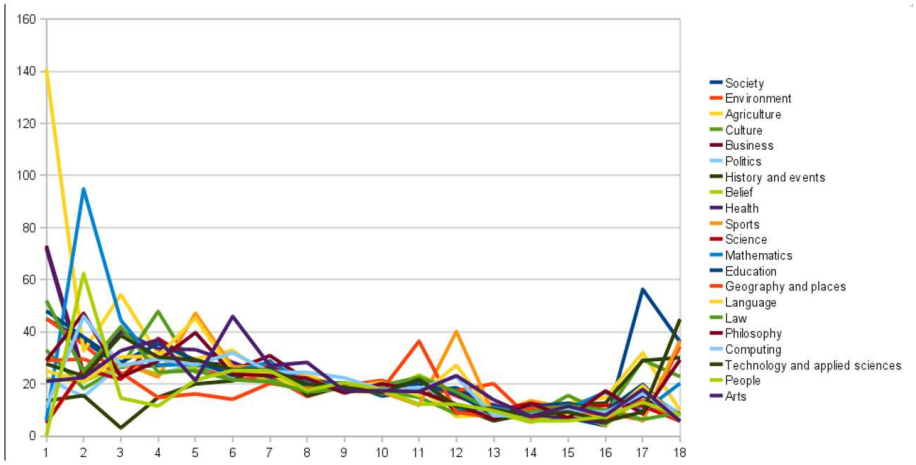

Figura 5.9:
La media degli articoli contenuti in ogni categoria al variare della distanza dalle macrocategorie.

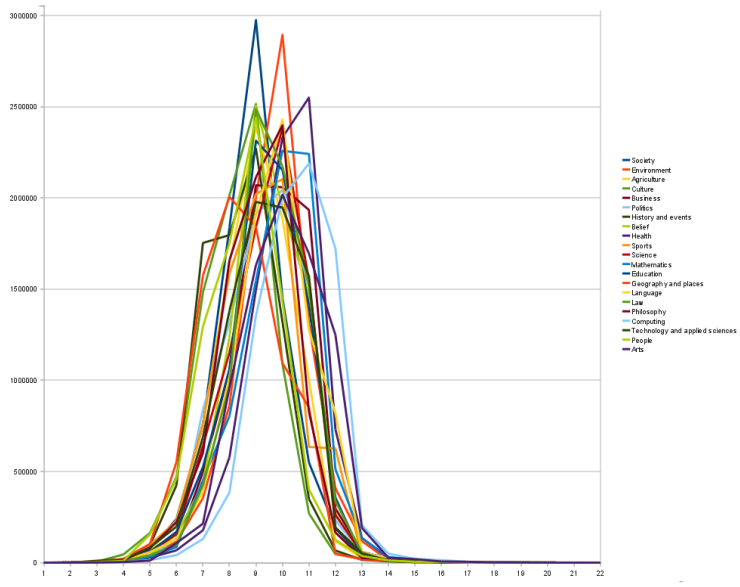

Figura 5.10:
La distribuzione di frequenza delle distanze delle categorie dalle macrocategorie, con un peso pari al numero di articoli contenuti.

profondità tassonomica. La curva di *Agriculture* è spostata verso sinistra ed è una delle prime a salire e ad avere un picco, per poi scendere molto velocemente. La macrocategoria *Agriculture*, infatti, non ha nessuna categoria a distanze maggiori di 17 passi, mentre le altre macrocategorie hanno dei nodi categoria con un percorso minimo fino a 21 passi.

Si nota che le curve sono molto simili tra di loro. Questo è causato dal fatto che, seguendo gli archi in entrambe le direzioni, si giunge facilmente da un nodo di una macrocategoria a un altro tramite delle categorie contenute in entrambi. Di conseguenza ogni curva è pesantemente influenzata dalle altre, perché nell'esplorazione del grafo a partire da una macrocategoria si giunge velocemente a esplorare le altre macrocategorie.

Una volta calcolate le appartenenze degli articoli basandosi sul nuovo grafo si nota che la differenza più importante rispetto ai risultati ottenuti senza normalizzazione è una diminuzione dei casi di sovrapposizione negli assegnamenti.

Questo cambiamento è dovuto al fatto che i baricentri sono in genere dei numeri non interi diversi tra di loro, quindi una categoria non sarà quasi mai equidistante da due macrocategorie come avveniva quando le proprietà *FROMx* erano degli interi, ma ci sarà quasi sempre una piccola differenza. Rimangono invece le sovrapposizioni dovute all'appartenenza di un articolo a più categorie assegnate a macrocategorie differenti tra loro.

Gli assegnamenti delle pagine d'esempio sono

```
Milan:History and events:100;
Italy:Culture:30;History and events:70;
Politecnico di Milano:Education:87,5;History and
    events:12,5;
Java (software platform):Computing:50;Technology and
    applied sciences:50;
Java:Geography and places:100; (L'isola principale
    dell'Indonesia)
Earth:Geography and places:100;
Albert Einstein>Society:1,79;People:5,36;Mathematics
    :92,86;
Nintendo:Agriculture:10;History and events:70;
    Philosophy:10;Technology and applied sciences:10;
20th Century Boys: Environment:46,87;Education:46,87;
    History and events:6,25;
Semiconductor fuse:Science:100;
```

Non ci sono state differenze sostanziali rispetto al caso base. Si nota

che le macrocategorie a cui sono stati assegnate le pagine *Albert Einstein, Nintendo* e *Semiconductor fuse* sono diminuite di numero, ma non c'è stato un miglioramento della correttezza.

La normalizzazione svolta tramite la moltiplicazione, infatti, ha soltanto fatto sì che le proprietà *FROMx* non siano quasi mai uguali tra di loro, diminuendo drasticamente i casi di categorie equidistanti da più macrocategorie, ma senza aggiungere molti assegnamenti corretti. Un approccio forse più corretto consiste nell'utilizzare la somma anziché la moltiplicazione per bilanciare i baricentri.

Questo metodo, come anticipato, non è stato utilizzato in questo caso perché la similitudine tra le curve lo renderebbe comunque inutile.

Il test di correttezza degli assegnamenti fornisce come risultato 0.20, quindi non si è ottenuto un miglioramento, anzi c'è stato addirittura un deciso peggioramento rispetto al valore di 0.34 ottenuto con il metodo di base. Il peggioramento potrebbe essere dovuto al fatto che gli utenti di Wikipedia tendono a suddividere una categoria in categorie più specifiche quando contiene troppi articoli, poiché una categoria troppo grossa è scomoda da utilizzare. Quindi, se esistono numerose pagine su un argomento, come accade per la geografia o per le persone, queste tenderanno ad essere organizzate in categorie molto specifiche e tassonomiche.In questo modo, poiché ogni categoria contiene un numero medio di sottocategorie piú alto, é possibile raggiungere piú categorie a partire da una macrocategoria in un numero prefissato di passi. Questo però non è un errore e quindi è scorretto allontanate dalla macrocategoria che è la più adatta a contenerle.

5.6 Percorso minimo nella direzione delle relazioni

Assegnando gli articoli alle macrocategorie solo con percorsi che seguano l'orientamento degli archi si ha una lista di soli 1836873 risultati, ossia il 65% dei 2822154 articoli categorizzati precedentemente utilizzando percorsi che ignoravano l'orientamento degli archi *SUBCATEGORYOF*.

Si possono esaminare gli assegnamenti delle pagine di esempio:

```
Milan:History and events:100;
Italy:Education:12,5;History and events:62,5;Belief:12,5;People:12,5;
Politecnico di Milano:History and events:75;Geography and places:25;
Java (software platform):Computing:100;
Java:Geography and places:100;
Earth:Geography and places:100;
Albert Einstein:Society:69,09;Science:12,73;People:18,18;
```

```
Nintendo:Agriculture:10;Culture:40;Computing:50;
20th Century Boys:Culture:100;
```

Semiconductor fuse è assente perché irraggiungibile dalle macrocategorie con passaggi lungo l'orientamento degli archi.

Mentre la pagina *Milan* è rimasta uguale, *Italy* viene assegnata erroneamente a *Education*, *Belief* e *People* perdendo *Culture.*

La pagina *Politecnico di Milano* perde l'assegnamento a *Education* ma viene assegnata a *Geography and places.*

L'articolo sul linguaggio Java viene attribuito completamente a *Computing,* perdendo la quota del 50% di *Technology and applied sciences*, e questo può essere considerato un miglioramento, mentre l'articolo sull'isola di *Java* rimane inalterato così come *Earth.*

La pagina *Nintendo* subisce un forte cambiamento: non è più assegnata per la maggior parte a *History and events*, che sparisce dai suoi assegnamenti, ma a solo tre macrocategorie tra cui *Computing* che detiene la quota più grossa. Rimane ancora l'anomalia di *Agriculture* ma c'è stato comunque un aumento della precisione.

Anche *20th Century Boys* riceve un assegnamento più plausibile, senza essere più assegnato a *Sports* o *History and events* ma solo a *Culture*, categoria in cui ha senso collocare un fumetto.

La voce *Albert Einstein* viene assegnata a tre categorie, *Society*, *Science* e *People*, che la definiscono in maniera più simile alla ripartizione che farebbe un valutatore umano rispetto al caso base in cui era ripartita in tante macrocategorie.

Utilizzando il metodo di misurazione della correttezza visto prima, basato sul confronto fra gli assegnamenti fatti da una persona e quelli fatti dall'algoritmo, si calcola una precisione di 0.35 dei nuovi risultati.

C'è quindi stato un miglioramento rispetto al 0.34 ottenuto senza questa variante, anche se assolutamente non decisivo soprattutto se si considera che è stata classificata poco più della metà degli articoli totali.

Assumendo che la correttezza media nel caso base non cambi se si considera solo il sottoinsieme dei nodi raggiungibili tramite gli archi orientati da almeno una macrocategoria non e facendo una media pesata tra i due risultati si calcola facilmente che la precisione dei risultati ottenuti utilizzando percorsi orientati, quando ne esiste almeno uno, o se non ci sono utilizzando anche percorsi non orientati, è $0,35*0,65 + 0,35*0,34 = 0.3465$.

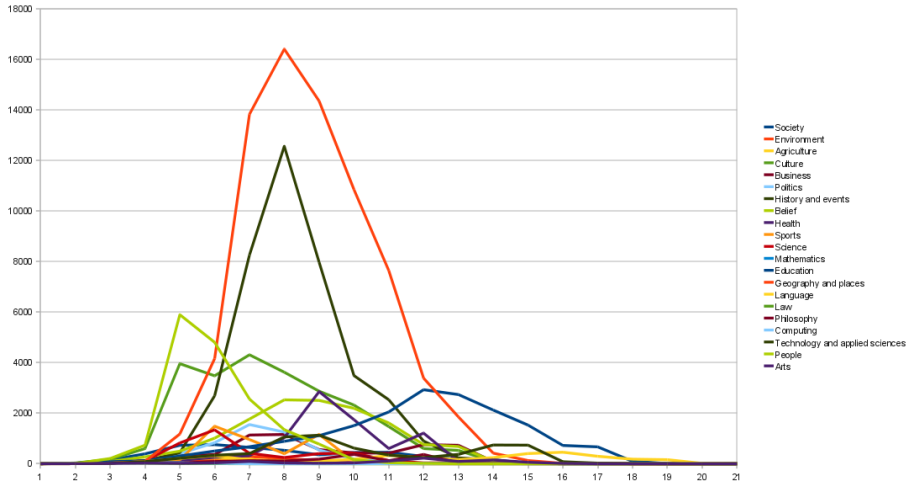

Figura 5.11: La distribuzione di frequenza delle distanze delle categorie dalle macrocategorie, a cui è stato assegnato un peso unitario, nel caso dei percorsi lungo l'orientamento degli archi.Le categorie *Geography and places* e *History and events* hanno dei picchi nettamente superiori a quelli delle altre macrocategorie.

5.7 Spostamento dei baricentri con percorsi diretti

Dopo avere assegnato le proprietà FROMx si calcolano le curve della distribuzione di frequenza delle distanze topologiche, ovviamente con percorsi che tengono conto degli orientamenti degli archi, ottenendo dei risultati (Figura 5.11) che mostrano la predominanza delle macrocategorie *Geography and places* e *History and events*. Anche *Culture* e *People* sono oggetto di numerosi assegnamenti di proprietà FROMx.

Si nota che le curve non sono più molto simili fra di loro come nel caso dei percorsi in direzione contraria all'orientamento degli archi. Questo accade perché non si può passare facilmente tra i nodi rappresentanti le sottocategorie di due diverse macrocategorie, e quando questo accade non si possono risalire gli archi per esplorare il ramo della macrocategoria raggiunta, come accadeva nel caso precedente.

I baricentri ottenuti sono

Language	12.25
Education	11.78
Arts	11.01
Technology and applied sciences	10.27
Philosophy	9.49
Health	9.38
Business	9.06
Belief	8.86
Geography and places	8.8
History and events	8.32
Society	7.8
Culture	7.63
Politics	7.22
Mathematics	7.21
Sports	7.17
Sciences	6.99
Environment	6.64
People	6.03
Agriculture	5.34
Computing	4.06
Law	3.29

Gli assegnamenti delle pagine di esempio sono:

```
Milan:History and events:100;
Italy:History and events:100;
Politecnico di Milano:History and events:50;Geography
    and places:50;
Java (software platform):Computing:100;
Java:Geography and places:100;
Earth:Geography and places:100;
Albert Einstein:Society:15,09;History and events
    :20,75;Science:62,26;People:1,89;
Nintendo:Culture:60;Technology and applied sciences
    :40;
20th Century Boys:Culture:100;
```

Semiconductor fuse è ovviamente assente così come era assente prima della normalizzazione.

Come sempre si calcola la dimensione delle macrocategorie (Figura 5.12).

Si nota, rispetto al caso base, un netto aumento delle macrocategorie

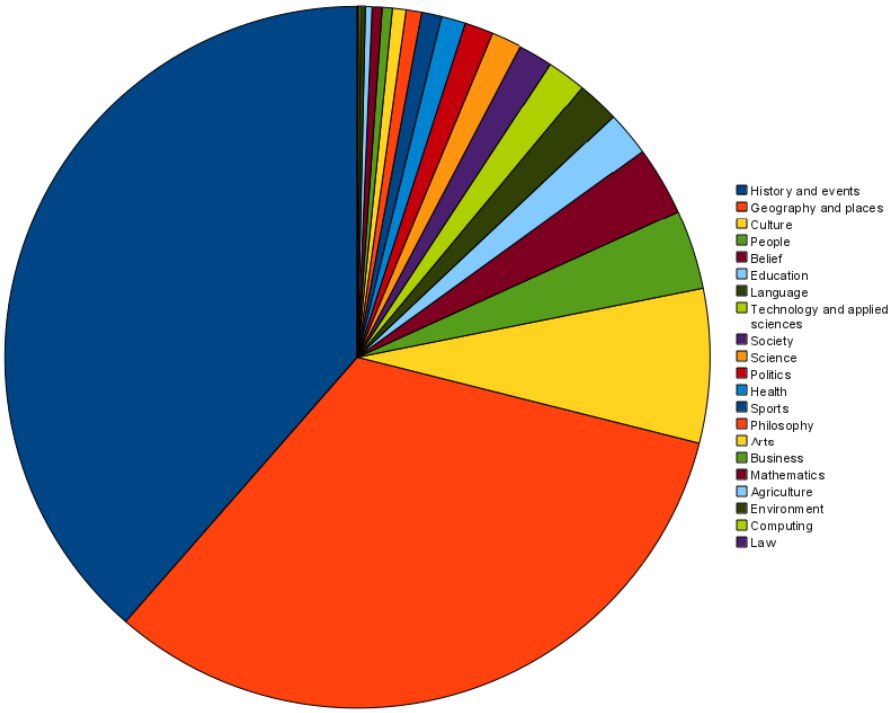

Figura 5.12:
Le dimensioni delle macrocategorie, in termini di quote totali di pagine assegnate, nel caso del grafo normalizzato per sottrazione con i percorsi dalle macrocategorie lungo la dimensione.

History and events e *Geography and places*, così come *Belief*, che diventa la quinta in ordine di dimensione mentre nel caso base era la diciassettesima.

Agriculture subisce un notevole ridimensionamento, passando dall'essere la quinta più grande nel caso base ad essere una delle più piccole, come era prevedibile viste le curve di distribuzione delle frequenze delle distanze dalle macrocategorie.

Effettuato il confronto con gli assegnamenti svolti manualmente dall'utente si ottiene un coefficiente di correttezza di 0.32.

Dunque la normalizzazione calcolata sulle distanze basate sugli archi orientati e effettuata tramite sottrazione e non moltiplicazione è più efficace, ma comunque sconveniente, in termini di precisione degli assegnamenti, rispetto ai risultati provenienti dal grafo non normalizzato, considerato anche che questa tecnica assegna il 65% degli articoli.

5.8 Costo di attraversamento differenziato in base alla direzione di orientamento degli archi

Effettuati gli assegnamenti si nota, scorrendo il file dei risultati, che molti articoli sono classificati in un numero minore di macrocategorie rispetto all'algoritmo originale, ossia c'è stata una diminuzione delle sovrapposizioni.

Questa è causata da una minore varianza nella gaussiana che approssima le curve di distribuzione delle categorie (Figura 5.13), dando a ognuna un peso uguale al numero di articoli contenuti.

Graficamente, la curva appare meno alta e più larga di quella ottenuta con i costi sempre unitari, segno di una distribuzione più omogenea delle categorie tra le distanze possibili.

Inoltre, mentre prima la distanza massima di una categoria era 21 e già a una distanza di 14 passi le categorie diventavano pochissime, ora ci sono categorie anche a una distanza di 25 passi.

Un esempio fra i tanti è l'articolo sull'Austria, la cui classificazione passa da

```
Austria>Society:18,33;Culture:18,33;Law:8,33;
    Philosophy:18,33;People:18,33;Science:18,33;
```

a

```
Austria>Law:25;People:75;
```

L'assegnamento anomalo a *People* è causato dai passaggi

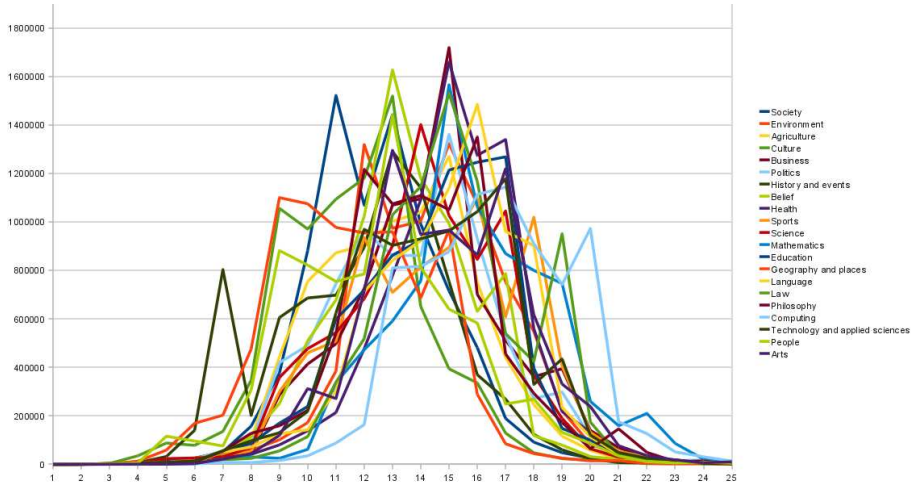

Figura 5.13: La distribuzione di frequenza delle distanze delle categorie dalle macrocategorie, con un peso pari al numero di articoli contenuti, nel caso dell'assegnamento delle distanze topologiche dando al percorso un costo dipendente dalla direzione di attraversamento degli archi.

```
Austria -> Erasmus Prize winners -> Award winners ->
    People by status -> Whistleblowers -> People by
    behavior -> People
```

mentre a *Law* si arriva con il percorso

```
Austria -> Federal countries -> Federalism ->
    Political theories -> Pacifism -> Core issues in
    ethics -> Law
```

Si possono contare gli articoli assegnati contemporaneamente a un certo numero di macrocategorie, ottenendo questi risultati

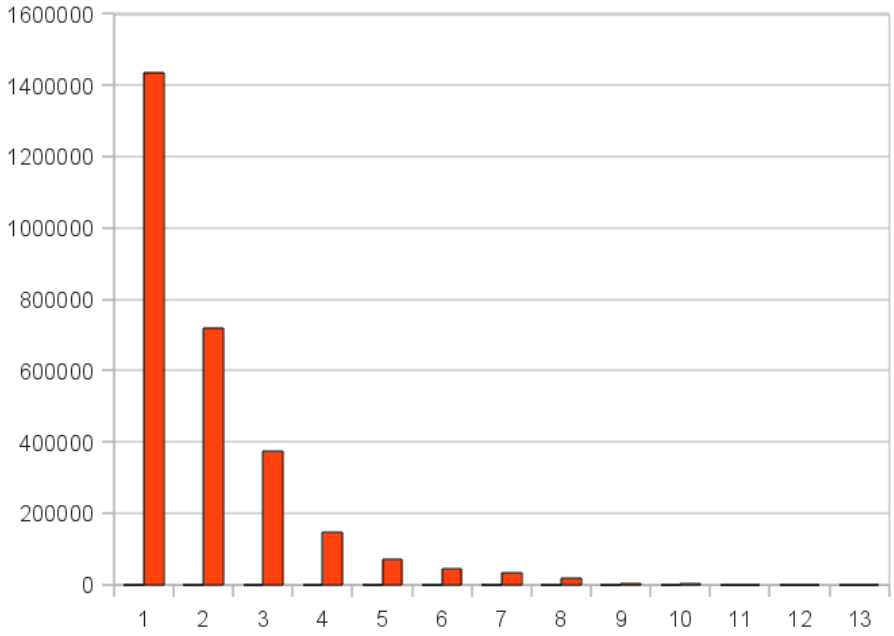

Figura 5.14: Il numero di articoli assegnati contemporaneamente a un certo numero di macrocategorie.

1	1436369
2	719570
3	372885
4	145588
5	69393
6	42805
7	32011
8	17143
9	2820
10	1039
11	139
12	23
13	5

La curva delle distribuzioni cresce molto velocemente (Figura 5.14) e gli articoli assegnati esattamente a una macrocategoria sono da soli il 51% del totale.

Questo risultato è conforme a quanto detto prima sulla distribuzione più ampia delle proprietà FROMx

La dimensione delle macrocategorie è rappresentata in Figura 5.15

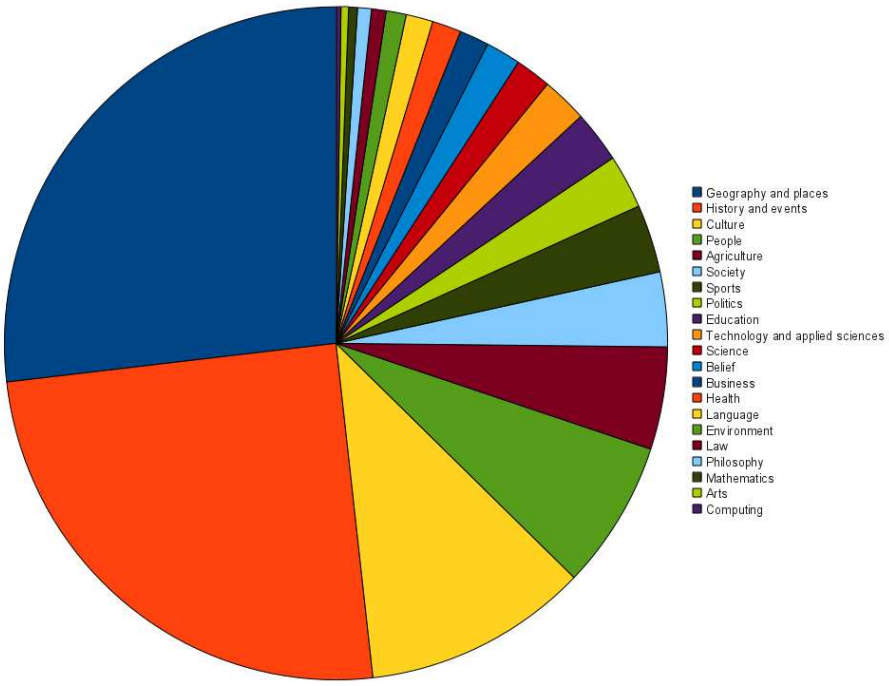

Figura 5.15:
La dimensione delle macrocategorie ottenute con il criterio dei costi differenziati. Ad ogni articolo, come per gli altri criteri, è assegnato un peso pari alla percentuale di assegnamento alla macrocategoria esaminata.

si nota che la macrocategoria *Geography and places* si è ingrandita notevolmente, ma nel complesso non ci sono differenze statistiche significative rispetto al caso base.

Si possono osservare gli assegnamenti delle pagine d'esempio

```
Milan:History and events:100;
Italy:Culture:30;History and events:70;
Politecnico di Milano:Education:87,5;History and
    events:12,5;
Java (software platform):Computing:50;Technology and
    applied sciences:50;
Java:Geography and places:100;
Earth:Geography and places:100;
Albert Einstein:Society:5,36;History and events:10,12;
    Science:39,58;People:44,94;
Nintendo:Agriculture:7,5;Culture:7,5;History and
    events:70;Geography and places:7,5;Philosophy:7,5;
20th Century Boys:Society:7,14;Environment:7,14;
    Culture:57,14;Education:7,14;Sports:7,14;Science
    :7,14;Technology and applied sciences:7,14;
Semiconductor fuse:Business:50;Science:50;
```

L'assegnamento delle pagine *Milan, Italy, Politecnico di Milano, Java (software platform), Java* e *Earth* rimane invariato rispetto al caso base, come *Semiconductor fuse*.

La voce su Albert Einstein viene assegnata a un numero minore di macrocategorie, ma quale dei due assegnamenti sia il più corretto è difficile da stabilire perché molto arbitrario.

L'assegnamento di *Nintendo* cambia ma rimane sbagliato e perde anche la quota di 1,67 destinata a *Technology and applied sciences.*

La voce sul fumetto *20th century boys* viene assegnata praticamente come nel caso base, salvo delle piccole variazioni dei valori delle quote.

Effettuando la valutazione manuale della correttezza si ottiene un valore di 0.37, quindi c'è stato un miglioramento rispetto al caso base e anche rispetto all'assegnamento svolto seguendo gli archi solo lungo la loro direzione, considerando che questo metodo classifica tutte le pagine collegate al grafo e non solo il 65% come nel caso degli archi orientati.

5.9 Costo di attraversamento degli archi basato sulle proprietà locali del grafo

Si è visto che è conveniente considerare la direzione di percorrenza degli archi, come criterio per decidere il costo di attraversamento, nella ricerca del percorso minimo. Tuttavia rimangono numerosi errori di assegnamento, come la pagina *Achilles* che viene assegnata a *Agriculture* al 20%

```
Achilles>Agriculture:20;Culture:40;People:40;
```

Dunque serve un'euristica che permetta di identificare quali tipi di archi vanno penalizzati perché portano frequentemente a degli errori.

Si nota che esistono delle categorie molto generiche, che si possono riconoscere per la grande quantità di categorie che contengono direttamente, quindi si vorrebbe penalizzarle.

Dunque si adotta come criterio per valutare il costo di attraversamento di un arco la quantità di categorie contenute in quella rappresentata dal nodo di arrivo.

Esistono numerosi modi di applicare questo criterio: si potrebbe assegnare un costo c fisso incrementato di un valore pari a un coefficiente p moltiplicato per il numero di categorie contenute, oppure si potrebbe utilizzare il logaritmo di tale numero per considerare l'ordine di grandezza più che la dimensione effettiva. Si è scelto per semplicità di dare un costo pari direttamente al numero di categorie contenute nella categoria di arrivo dell'arco.

Una volta effettuato il calcolo delle distanze di procede con l'assegnamento e con la valutazione dei risultati ottenendo un coefficiente di correttezza di 0.17, quindi decisamente basso rispetto ai valori ottenuti con gli altri metodi.

Si potrebbero utilizzare altre metriche come quella logaritmica, andando a tentativi, ma per provare ognuna delle tecniche servirebbero molte ore di calcolo, nello specifico sono stati necessari circa due giorni per effettuare il calcolo delle distanze e l'assegnamento degli articoli, e in presenza di un coefficiente di correttezza così basso appare chiaro che non è una soluzione che conviene esplorare.

Osservando allora le anomalie negli assegnamenti, come nel caso della voce Austria citata prima, delle pagine come *RAI*, classificata completamente in *Geography and places*, o *Atomic physics*, classificata totalmente in *Belief*, si possono cercare delle caratteristiche comuni ai percorsi minimi che portano agli errori.

```
Austria -> Erasmus Prize winners -> Award winners ->
    People by status -> Whistleblowers -> People by
    behavior -> People
RAI -> Government-owned companies in Italy ->
    Government-owned companies by country -> Government
     by country -> Country subdivisions -> Geography by
     place -> Geography -> Geography and places
Atomic physics -> Atomic, molecular, and optical
    physics -> Physics -> Reality -> Alternate reality
    -> Belief
```

Il terzo caso di assegnamento è causato dall'aver percorso l'arco tra *Reality* e *Alternate Reality* contro il suo orientamento, ma si è visto che questo tipo di errore è difficilmente risolvibile poiché escludendo o penalizzando i percorsi che non seguono l'orientamento degli archi si ha un peggioramento della precisione dell'assegnamento.

I primi due errori, invece, sono riconducibili all'esistenza di tre categorie, *People by status*, *Government-owned companies by country* e *Government by country*, che hanno per loro natura degli archi con nodi rappresentanti categorie di aree semantiche diverse e portano quindi ad accorciare le distanze topologiche tra queste categorie.

Purtroppo non è possibile, allo stato attuale, identificare automaticamente una categoria generica come queste.

Si possono però trovare delle euristiche che identifichino una parte il più grande possibile di queste categorie e forniscano il minor numero possibile di falsi positivi.

Un'osservazione molto semplice che si può fare è che tutte e tre le categorie contengono la parola by nel nome.

Scorrendo la lista delle categorie che contengono la stringa " *by* ", spazi compresi, si nota che in effetti la maggior parte di queste tendono a collegare elementi che hanno buone probabilità di appartenere a aree semantiche diverse, come *1090s by country, 10th century by continent* o *19th-century people by nationality*.

In generale, ci sono categorie che organizzano le persone e gli eventi in base all'anno e alla nazione.

Si può quindi modificare l'algoritmo di assegnamento dei costi del caso base per aggiungere 1 al costo di percorrenza di tutti gli archi che portano a una categoria il cui nome contiene la sotto-stringa " by ".

Una volta assegnate le proprietà *FROMx* con questo criterio si effettuano gli assegnamenti e si valutano con il solito metodo.

Il risultato è 0.22, quindi ancora una volta non si è ottenuto un miglioramento bensì un peggioramento.

Le pagine prese ad esempio sono state classificate in questo modo:

```
Austria>Society:22,5;Culture:22,5;Law:10;Philosophy
    :22,5;Science:22,5;
RAI>Society:12,5;History and events:12,5;Belief:37,5;
    Sports:37,5;
Atomic physics>History and events:33,33;Belief:33,33;
    Geography and places:33,33;
Achilles:Law:20;Arts:80;
```

Quindi, per effetto della modifica, gli assegnamenti sbagliati sono effettivamente cambiati ma i nuovi assegnamenti sono comunque sbagliati.

5.10 Assegnamento maggioritario

Il file degli assegnamenti a una sola macrocategoria contiene 1840604 elementi, quello degli assegnamenti a più macrocategorie ne contiene 981550.

È possibile contare la quantità di articoli assegnati a ciascuna macrocategoria

Geography and places	494354
History and events	461559
People	204220
Culture	145481
Sports	122883
Agriculture	111895
Politics	56484
Education	51169
Society	46971
Technology and applied sciences	37771
Law	26623
Environment	17532
Language	11778
Business	10918
Computing	8831
Arts	8160
Belief	6871
Mathematics	6034
Health	4916
Philosophy	3707
Science	2447

e rappresentare questi dati graficamente (Figura 5.16). Si notano delle differenze rispetto al calcolo delle dimensioni delle macrocategorie basato sulla somma delle percentuali di assegnamento illustrato nel capitolo 1.

È possibile calcolare il cambiamento della quantità di assegnamenti ricevuti in rapporto al totale nei due casi.

La quota di assegnamenti alla macrocategoria *Geography and places* è aumentata del 26% rispetto alle sue dimensioni misurate con il metodo probabilistico, *History and events* del 23%, *Sports* del 20%, *People* dell'11%.

Le categorie che ricevono minori assegnamenti, in proporzione al totale, sono Science, che riceve solo l' 11% di quanto riceveva prima (infatti è quella che prende la minore quota mentre prima ce ne erano 7 più piccole) e *Philosophy*, che riceve il 31% rispetto a prima.

Questo cambiamento è dovuto alla tendenza di alcune macrocategorie a apparire frequentemente ma con piccole percentuali negli assegnamenti del caso base, risultando molto piccole nell'assegnamento maggioritario che ignora le piccole quote di appartenenza.

Si può effettuare la valutazione della correttezza degli assegnamenti, che in questo caso, per ogni articolo esaminato, restituirà un valore di 1 se la

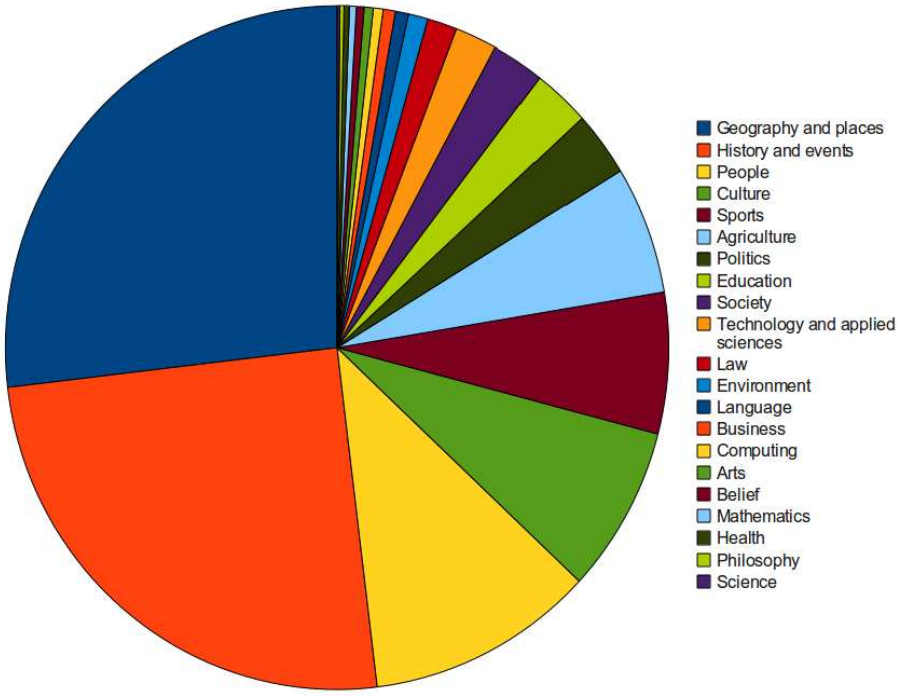

Figura 5.16:
Il grafico a torta mostra i risultati dell'assegnamento delle pagine alle macrocategorie con il criterio maggioritario.

macrocategoria scelta dall'utente è la stessa scelta dal programma e 0 se sono diverse.

Su 50 articoli esaminati, 14 sono stati assegnati alla stessa macrocategoria dall'algoritmo e dal valutatore, quindi la precisione è di 0.28. Questo dato non è confrontabile con quelli ottenuti con i precedenti metodi perché si sta valutando un altro tipo di misurazione.

Un'altra analisi interessante è quella della sovrapposizione delle macrocategorie.

Si hanno 3675 diversi set di macrocategorie che contengono una pagina. Di questi, 848 contengono solo una pagina, 324 ne contengono due e 221 ne contengono 3.

Ecco i set di macrocategorie più frequenti:

Culture,History and events	57439
Culture,People	50068
Agriculture,Culture	41025
Culture,History and events,People	36854
Geography and places,History and events	32092
Culture,Geography and places	30464
Mathematics,Science,Technology and applied sciences	18377

La macrocategoria con la maggiore tendenza a sovrapporsi alle altre è decisamente *Culture*, seguita da *History and events.* Si osserva che queste affinità non corrispondono totalmente a quelle ottenute applicando la cosine similarity agli assegnamenti percentuali, per esempio la coppia *Mathematics* e *Science* non è più quella più diffusa, così come quella formata da *Society* e *Health.*

Anche se gli abbinamenti sono ottenuti in maniera diversa (nel caso base con la cosine similarity e in questo caso con la frequenza con cui compaiono entrambi gli assegnamenti), la causa di questa differenza è dovuta al fatto che l'assegnamento maggioritario, come si è già visto, non tiene conto delle macrocategorie che ricevono gli assegnamenti di piccole quote di molti articoli.

5.11 Assegnamento con ripartizione di punteggi

Vengono assegnati 1757593 articoli, ossia il 62% di quelli connessi al grafo e elaborati con il metodo base. La differenza rispetto al 65% ottenuto con l'assegnamento ottenuto percorrendo gli archi solo nella loro direzione è dovuta all'eliminazione dei cicli.

Infatti, eliminando i cicli si rendono irraggiungibili alcune categorie che

non sono collegate al reticolo dei percorsi diretti ottenuto se non tramite le relazioni che sono state eliminate per non avere cicli.

Se queste categorie diventano irraggiungibili lo diventano anche quelle che erano raggiungibili solo attraverso di loro, questo è il motivo per cui si perde il 3% delle categorie.

Le pagine di esempio vengono categorizzate in questo modo:

```
Milan:History and events:100;
Italy>History and events:100;
Politecnico di Milano:History and events:7,85;
    Geography and places:92,15;
Java (software platform):Computing:100;
Java:Geography and places:100;
Earth:Geography and places:100;
Albert Einstein:Society:28,19;Culture:0,21;Education
    :0;History and events:12,14;Belief:13,4;Geography
    and places:0;Science:40,79;People:5,27;
Nintendo:Culture:4,27;Business:0,46;Politics:5,96;
    History and events:89,09;Geography and places:0;
    Technology and applied sciences:0,23;
20th Century Boys:Culture:94,96;History and events
    :5,04;
```

La pagina *Semiconductor fuse* non è stata assegnata perché non è raggiungibile da nessuna macrocategoria con passaggi nella direzione degli archi.

La pagina *Politecnico di Milano* è stata assegnata con minore precisione, senza nessuna quota assegnata a *Education*.

L'assegnamento della pagina sul linguaggio Java solo a *Computing* (e non anche per il 50% a *Technology and applied science*) si può invece considerare un miglioramento, così come *20th Century Boys* che riceve un assegnamento quasi totale a *Culture* anziché essere assegnato a numerose macrocategorie chiaramente sbagliate come *Environment*.

La pagina *Nintendo* viene assegnata in una maniera diversa rispetto a prima, ma comunque molto imprecisa, infatti riceve una quota molto larga di *History and events* e una ancora più bassa di *Technology and applied sciences*.

La pagina su Albert Einstein non subisce miglioramenti o peggioramenti essenziali, la quota assegnata a People scende da 24.78 a 5.27 ma *Science* sale da 18.53 a 40.79, dunque alcune quote diventano più simili a quelle che si assegnerebbero a mano e altre si allontanano dal valore corretto.

Osservando la dimensione delle macrocategorie (Figura 5.17), ottenu-

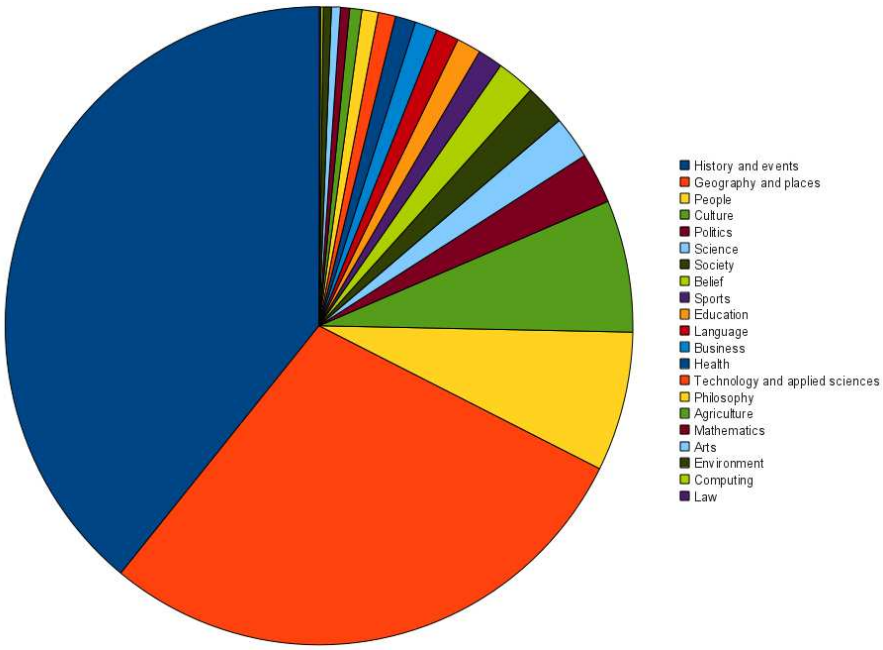

Figura 5.17:
La dimensione delle macrocategorie, in termini di numero di assegnamenti ricevuti, ottenute con il metodo della ripartizione del punteggio.

ta come sempre contando il numero di articoli contenuti, si osserva che la macrocategoria *History and events* si è notevolmente espansa, mentre *Agriculture* è diventata una delle più piccole.

Questo risultato indica che, nel sotto-grafo degli archi e dei nodi raggiungibili dalla macrocategoria *Agriculture*, si ha una scarsa quantità di archi, e di percorsi diversi.

History and events, invece, ha una struttura molto più complessa, ricca di percorsi alternativi per raggiungere gli stessi nodi a partire dalla radice.

Il conteggio del numero di articoli assegnati a una certa quantità di macrocategorie restituisce questi risultati:

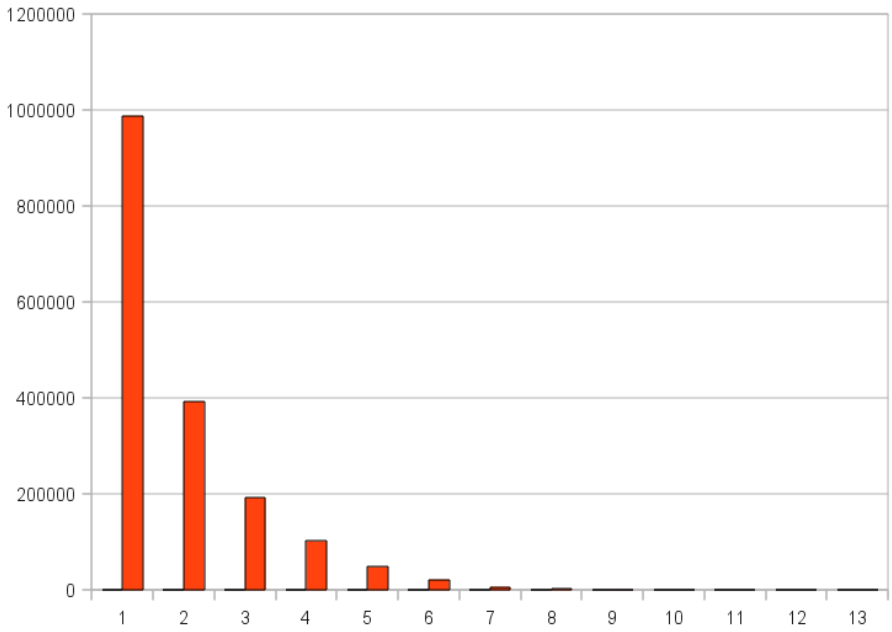

Figura 5.18: La quantità di articoli assegnati a un certo numero di macrocategorie. La curva decresce velocemente e regolarmente.

1	988335
2	393133
3	192855
4	102292
5	48648
6	22258
7	6836
8	2630
9	479
10	100
11	29
12	1
13	1

In questo caso non ci sono anomalie, la curva (Figura 5.18) è monotona decrescente.

La maggior parte degli articoli, il 79%, risulta assegnata a una o due macrocategorie.

L'articolo assegnato a 12 macrocategorie è *Wayne Gretzky*, una voce sul più importante giocatore di hockey su ghiaccio canadese, che è assegnato in

questo modo:

```
Wayne Gretzky>Society:7,44;Education:0;Culture:0,02;
    Business:0,12;Politics:55,12;History and events
    :0,13;Geography and places:0;Belief:0,05;Health
    :0,05;Sports:36,08;People:0,93;Technology and
    applied sciences:0,06;
```

inoltre la pagina è contenuta direttamente in ben 46 categorie, un numero molto più alto del solito, dovuto probabilmente all'attenzione degli editori alla pagina dovuta alla gran fama del personaggio.

La voce assegnata a 13 macrocategorie è invece quella su *George Steinbrenner*, ex proprietario dei *New York Yankees*, che viene assegnato a queste macrocategorie:

```
George Steinbrenner>Culture:0;Business:0;Politics
    :0,04;History and events:0;Belief:0;Health:0,13;
    Sports:99,43;Education:0;Geography and places:0;
    Language:0;Philosophy:0,01;Technology and applied
    sciences:0;People:0,38;
```

anche questo articolo appartiene a molte categorie, 17, che pur essendo un valore alto rappresenta un caso abbastanza comune.

Si osserva che in entrambi i casi ci sono degli assegnamenti pari a 0. Questa è un'anomalia dovuta al troncamento, che fa sì che un'appartenenza molto bassa, con un valore inferiore a 0.005, appaia con un valore di 0.

Questo però non influisce sul calcolo delle dimensioni delle macrocategorie, perché un assegnamento così basso non influenza la statistica.

La precisione degli assegnamenti risulta essere 0.35, quindi la qualità dei risultati è praticamente la stessa del caso dell'assegnamento basato sulla semplice distanza topologica con percorsi orientati, sia come numero di pagine assegnate (c'è stata una variazione in peggio del 3%) che come qualità degli assegnamenti.

5.12 Assegnamento con probabilità di raggiungere la macrocategoria

Vengono assegnati anche in questo caso 1757593 articoli, ossia il 62% di quelli connessi al grafo e elaborati con il metodo base.

Gli articoli di esempio vengono classificati nel seguente modo:

```
Milan:History and events:100;
```

```
Italy>History and events:100;
Politecnico di Milano:History and events:86,53;
    Geography and places:13,47;
Java (software platform):Computing:100;
Java:Geography and places:100;
Earth:Geography and places:100;
Albert Einstein:Society:12,37;Culture:1,18;Education
    :0,09;History and events:3,52;Belief:0,16;Geography
     and places:0,31;Science:22,94;People:59,41;
Nintendo:Culture:16,5;Business:26,82;Politics:4,13;
    History and events:44,56;Geography and places:1,29;
    Technology and applied sciences:6,7;
20th Century Boys:Culture:45,45;History and events
    :54,55;
```

Rispetto a prima si osserva che è migliorato l'assegnamento di *Albert Einstein*, che viene collegato soprattutto a *People* e *Science*.

Anche *Nintendo*, pur ricevendo una fetta di *History and events* che difficilmente verrebbe data da un valutatore umano, viene abbinata di più a *Business* e a *Culture* rispetto all'assegnamento con ripartizione di punteggi.

20th century boys viene assegnata a *History and events* con un punteggio maggiore rispetto all'assegnamento con ripartizione dei punteggi ma nel complesso è comunque un assegnamento più plausibile di quello del caso base.

La tabella delle dimensioni delle macrocategorie, in termini di quote di articoli assegnate, è

History and events	71391501
Geography and places	50150533
People	13001061
Culture	12915638
Belief	3779893
Science	3739630
Society	3322713
Business	2860546
Education	2638730
Language	1760599
Health	1712138
Politics	1611366
Sports	1476404
Technology and applied sciences	1382722
Philosophy	1056561
Agriculture	891377
Arts	715559
Environment	588917
Mathematics	446926
Computing	227235
Law	115038

e anche dalla rappresentazione grafica (Figura 5.19) si nota come anche in questo caso *Agriculture* sia penalizzata, ottenendo una fetta piccola come ci si aspetterebbe, e gli argomenti di storia e geografia costituiscano insieme la maggior parte, precisamente il 69%, dei temi trattati dall'enciclopedia.

Il conteggio del numero di macrocategorie abbinate a ciascun articolo fornisce i seguenti risultati:

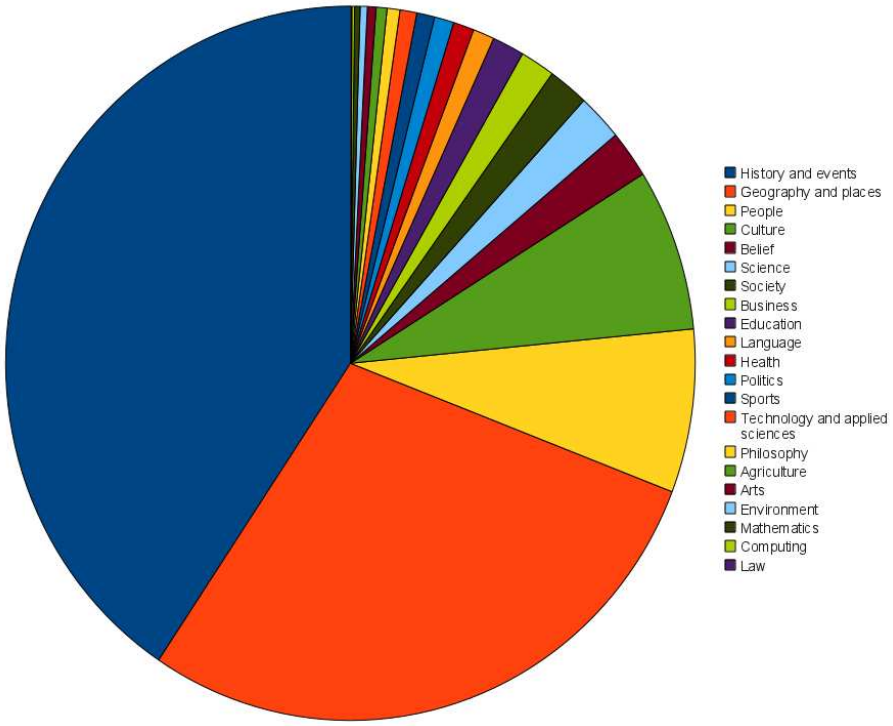

Figura 5.19:
La dimensione delle macrocategorie nel caso degli assegnamenti con probabilità di raggiungimento della macrocategoria a partire dall'articolo muovendosi lungo l'orientamento degli archi.

1	987581
2	392966
3	193600
4	102668
5	48686
6	22269
7	6840
8	2631
9	480
10	100
11	29
12	1
13	1

Si osserva (Figura 5.20) che la curva è monotona decrescente e diminuisce velocemente.

La maggior parte degli articolo, il 56%, viene abbinata a una macrocategoria, e il 79% a meno di tre.

I due articoli abbinati a 12 e 13 macrocategorie sono, rispettivamente, *Wayne Gretzky* e *George Steinbrenner*, gli stessi di prima.

Gli assegnamenti sono:

```
George Steinbrenner:Culture:0,18;Business:18,96;
    Politics:3,36;History and events:6,91;Belief:0,18;
    Health:6,34;Sports:31,68;Education:0,05;Geography
    and places:1,32;Language:0,12;Philosophy:0,72;
    Technology and applied sciences:4,74;People:25,44;
Wayne Gretzky:Society:53,03;Education:0,07;Culture
    :1,02;Business:3,57;Politics:5,1;History and events
    :1,76;Geography and places:0,09;Belief:1,02;Health
    :1,1;Sports:5,48;People:26,86;Technology and
    applied sciences:0,89;
```

La valutazione della correttezza degli assegnamenti restituisce un valore di 0.36. Si è dunque avuto un miglioramento sia rispetto al caso base che rispetto all'assegnamento con ripartizione di punteggi, ma non rispetto all'assegnamento con costi differenziati rispetto alla direzione di attraversamento degli archi, che era 0.37, considerando anche che viene assegnato il 62% degli articoli totali.

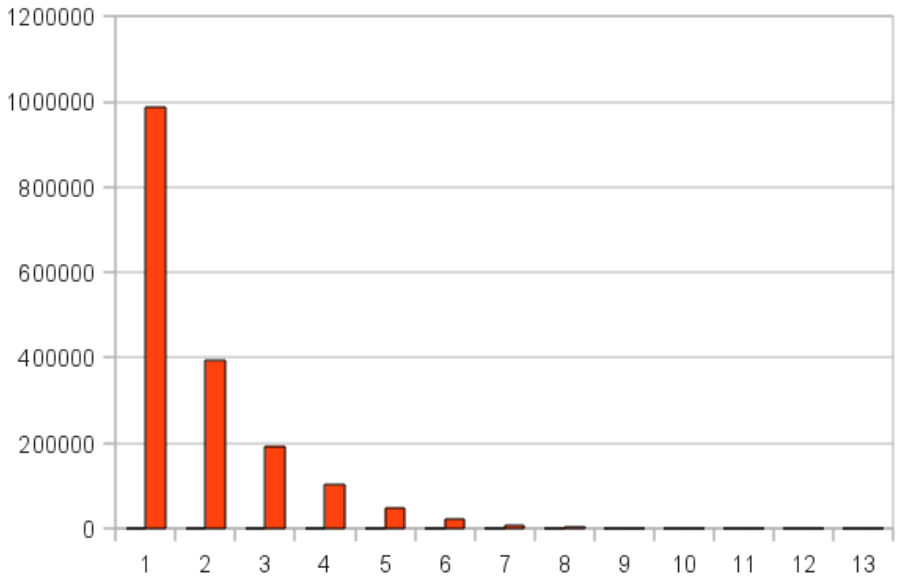

Figura 5.20: La quantità di articoli assegnati a un certo numero di macrocategorie con il criterio della probabilità di raggiungimento. La curva decresce velocemente e regolarmente.

5.13 Discussione dei risultati

I valori di correttezza degli assegnamenti ottenuti con i vari metodi sono:

Metodo	valore di correttezza	% pagine valutate
Caso base (Kittur, Holloway)	0.34	100%
Solo percorsi diretti	0.35	65%
Costi differenziati	0.37	100%
Costo dipendente dal numero di categorie contenute	0.17	100%
Normalizzazione per moltiplicazione	0.20	100%
Normalizzazione per sottrazione sui percorsi diretti	0.32	65%
Penalizzazione sottostringa " *by* "	0.22	100%
Ripartizione dei punteggi	0.35	62%
Probabilità di raggiungimento	0.36	62%

In base alle valutazioni sperimentali svolte il metodo migliore tra quelli esaminati è dunque quello dei costi differenziati in base alla direzione di attraversamento degli archi, che assegna tutti gli articoli con una precisione di 0.37.

La percentuale di pagine valutate è riferita alle pagine raggiungibili da almeno una macrocategoria seguendo gli archi indipendentemente dall'orientamento, nel grafo filtrato. Dunque non sono considerate le pagine di

disambigua e i redirects e sono escluse anche le pagine non categorizzate o assegnate a categorie a loro volta non categorizzate, situazione anomala e comunque molto rara.

Si nota inoltre che, oltre a un diverso coefficiente di correttezza, i metodi di assegnamento presentano delle differenze nella dimensione delle macrocategorie in termini di articoli abbinati. La macrocategoria *Agriculture*, che nel caso base era la quinta in ordine di dimensioni, diventa la sedicesima sia nel caso dell'assegnamento basato sulla ripartizione dei punteggi che in quello basato sulle probabilità di raggiungerla partendo dal nodo pagina e risalendo gli archi scegliendoli a caso. Utilizzando la funzione *random page* di MediaWiki si osserva che gli argomenti più trattati nelle pagine sono la storia, la geografia, le persone e le opere d'arte (specialmente album musicali, film e libri), quindi queste ultime due varianti dell'algoritmo forniscono una dimensione delle macrocategorie più verosimile in cui *Agriculture* non ha più una presenza eccessiva.

È stato dunque dimostrato empiricamente che il criterio di Kittur fornisce dei risultati plausibili anche in presenza di un numero di categorie, articoli e macrocategorie selezionate più alto. Si sono individuati inoltre dei metodi più precisi dell'assegnamento basato sulla distanza topologica.

Capitolo 6

Conclusioni e sviluppi futuri

6.1 Conclusioni

In questa tesi sono stati provati vari metodi alternativi a quello di Kittur per svolgere l'assegnamento degli articoli alle macrocategorie. Dopo aver effettuato l'assegnamento basato sulla distanza topologica, ossia il caso base, sono state calcolate delle statistiche sulla copertura degli argomenti di Wikipedia e sulla sovrapposizione delle macrocategorie. Osservando gli assegnamenti di alcuni articoli e i loro percorsi minimi verso le macrocategorie si è visto che gli errori sono causati dalla diversa struttura delle categorie che si occupano di diversi argomenti e dalla presenza di archi nel grafo degli assegnamenti che legano argomenti distanti pur senza essere sbagliati, e si sono provate diverse varianti per gestire queste particolarità.

Per affrontare la diversa struttura delle categorie si è usata la normalizzazione e si sono applicati dei criteri per il calcolo del costo di attraversamento basati sulle proprietà dei singoli nodi, come la presenza di " *by* " nel nome o il numero di categorie contenute, senza successo. Osservando invece i percorsi che seguono gli archi in entrambe le direzioni, lungo l'orientamento o al contrario, si è pensato di percorrere gli archi solo lungo il loro orientamento in modo da non poter passare da una categoria a una più specifica per poi risalire ad un' altra categoria che la contiene raggiungendo una macrocategoria sbagliata, come accade per la pagina *Nintendo*. Il difetto di questa tecnica è che non tutte le pagine sono raggiungibili dalle macrocategorie senza mai variare il verso di percorrenza degi archi. Inoltre, non sempre è sbagliato seguire gli archi in entrambi i sensi, poiché comunque collegano delle categorie che hanno dei legami semantici.

Si è dunque provato ad assegnare agli archi percorsi nella direzione opposta all'orientamento un costo di attraversamento più alto. In questo mo-

do i passaggi che possono generare anomalie è scoraggiato ma non impedito, con il risultato che tutte le pagine sono raggiungibili dalle macrocategorie. Il risultato è stato positivo, infatti il punteggio di similitudine tra gli assegnamenti automatici e quelli svolti da un valutatore umano è aumentato.

Sono state infine provate delle strategie di assegnamento che tenessero conto di più percorsi contemporaneamente, nell'ipotesi che se un arco può costituire un'anomalia nel calcolo delle distanze topologiche considerando più percorsi diversi contemporaneamente se ne limitano gli effetti negativi. Anche queste tecniche hanno dato risultati migliori della semplice distanza topologica, senza però superare il metodo dei costi di attraversamento differenziati.

6.2 Sviluppi futuri

Il grafo su cui si è lavorato teneva conto solo degli assegnamenti alle categorie, ma non dei contenuti delle pagine. Per aumentare la precisione potrebbe dunque essere molto utile analizzare anche i collegamenti ipertestuali tra le pagine, in maniera simile a quanto fatto da Zesch[22], per rafforzare la validità delle relazioni di appartenenza. Per esempio, se la pagina *Milan* è la pagina principale dell'omonima categoria è ragionevole rafforzare il valore degli archi che dalla categoria *Milan* vanno a pagine con collegamenti ipertestuali con l'articolo su Milano. Integrando il grafo delle categorie con il grafo degli assegnamenti si potrebbe ottenere una minore suscettibilità dell'algoritmo agli assegnamenti anomali e alle parti del grafo con una struttura molto più gerarchica o al contrario molto più interconnessa rispetto alla norma.

Un'altra limitazione ad alcune tecniche è stata l'impossibilità di collegare molti articoli, tra il 35% e il 38%, con un percorso che rispettasse l'orientamento degli archi, ad almeno una macrocategoria. Questo è anomalo, visto che le macrocategorie scelte dovrebbero coprire, o almeno toccare, tutti gli argomenti possibili, e influenza alcune tecniche che richiedono un grafo dove questo non accada se non in pochissimi casi isolati. Sarebbe interessante mettere a punto una tecnica per trasformare il grafo in una struttura dove ogni nodo sia raggiungibile da almeno una macrocategoria, se non tutte, muovendosi sempre nello stesso modo rispetto all'orientamento degli archi, ossia una struttura interamente rappresentabile con un diagramma di Hasse. In questo modo si potrebbero applicare criteri come quello della probabilità di raggiungimento, magari combinandolo con il conteggio dei collegamenti tra le pagine di diverse categorie per dare un diverso peso ad ogni arco,

per ripartire in maniera più intelligente i valori di probabilità favorendo le categorie vicine anche in base ai collegamenti tra le pagine contenute.

Un'altra osservazione che si può fare è sul criterio di scelta degli articoli da usare per la valutazione. Scegliendoli a caso fra tutti quelli esistenti si selezionano a volte delle bozze, i cosiddetti *stub*, ossia degli articoli molto brevi composti da poche righe e spesso categorizzati in maniera molto scarsa. In questa tesi questo aspetto è stato ignorato poiché si cerca un algoritmo che effettui l'assegnamento con la maggiore precisione possibile e gli stub influenzano in maniera molto lieve i risultati, inoltre le pagine appena create o molto scarne non sono categorizzate se non nelle categorie degli stub, escluse dal grafo durante il filtraggio e quindi dall'assegnamento. A seconda dello scopo che ci si prefigge nell'assegnamento degli articoli alle macrocategorie potrebbe però essere conveniente o meno ignorare gli articoli troppo brevi, oppure troppo recenti o classificati come *stub* dalla struttura a categorie stessa. In questo modo si impedirebbe a questi articoli incompleti di alterare le statistiche sugli assegnamenti e di lavorare solo su quelle pagine da cui è possibile estrarre delle informazioni veramente utili.

Bibliografia

[1] Algoritmo di tarjan per le componenti fortemente connesse, versione del 3 settembre 2010. http://it.wikipedia.org/w/index.php?title=Algoritmo_di_Tarjan_per_le_componenti_fortemente_connesse&oldid=34630372.

[2] igraph, controllato il 26 agosto 2010. http://igraph.sourceforge.net/.

[3] Jaccard index, da wikipedia, versione del 4 settembre 2010. http://en.wikipedia.org/w/index.php?title=Jaccard_index&oldid=379972752.

[4] Main topic classifications, versione del 26 agosto 2010. http://en.wikipedia.org/w/index.php?title=Category:Main_topic_classifications&oldid=379868845.

[5] The neo database. http://dist.neo4j.org/neo-technology-introduction.pdf.

[6] Neo4j homepage. http://neo4j.org.

[7] Pagina di wikipedia di suggestbot, versione del 4 settembre 2010. http://en.wikipedia.org/w/index.php?title=User:SuggestBot&oldid=378422052.

[8] R project, controllato il 26 agosto 2010. http://www.r-project.org/.

[9] Wiki di supporto a pajek. http://pajek.imfm.si/doku.php.

[10] Wikipedia:faq/categorization, versione del 4 settembre 2010. http://en.wikipedia.org/w/index.php?title=Wikipedia:FAQ/Categorization&oldid=381149961#State_of_the_Category_feature.

[11] autori vari. Voce di wikipedia su se stessa, versione del 4 settembre 2010. http://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=382591253.

[12] Loren Terveen Dan Cosley, Dan Frankowski and John Riedl. Suggestbot: Using intelligent task routing to help people find work in wikipedia. 2008.

[13] Gjergji Kasneci Fabian M. Suchanek and Gerhard Weikum. Yago: A large ontology from wikipedia and wordnet. 2008.

[14] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. 2007.

[15] Aniket Kittur, Ed H. Chi, and Bongwon Suh. What's in wikipedia? mapping topics and conflicts using socially annotated category structure. 2009.

[16] Olena Medelyan and Catherine Legg. Integrating cyc and wikipedia: Folksonomy meets rigorously defined common-sense. 2008.

[17] Simone Paolo Ponzetto and Michael Strube. Deriving a large scale taxonomy from wikipedia. 2007.

[18] Michael Robinson Samuel Sarjant, Catherine Legg and Olena Medelyan. all you can eat ontology-building: Feeding wikipedia to cyc. 2009.

[19] Wolfgang Nejdl Sergey Chernov, Tereza Iofciu and Xuan Zhou. Extracting semantic relationships between wikipedia categories. 2006.

[20] Miran Bozicevic Todd Holloway and Katy Borner. Analyzing and visualizing the semantic coverage of wikipedia and its authors. 2006.

[21] Tim Finin Zareen Syed and Anupam Joshi. Wikipedia as an ontology for describing documents. 2008.

[22] Torsten Zesch and Iryna Gurevych. Analysis of the wikipedia category graph for nlp applications. 2007.

Appendice A

Contenuto del DVD allegato

Il DVD allegato contiene:

- Una copia di questa stessa tesi, in formato pdf
- I risultati degli assegnamenti piú importanti, in testo semplice, nella cartella *assegnamenti*
- Il codice sorgente (Java) di tutti i programmi creati per svolgere il lavoro, nella cartella *sorgenti*
- I file che descrivono il grafo, prima del filtraggio, delle categorie e degli articoli di Wikipedia del 12 marzo 2010, in formato testuale, nella cartella *struttura*
- Il database di Neo4j ottenuto dopo il filtraggio, nella cartella *database*
- Gli elenchi delle pagine valutate manualmente e i risultati del confronto con gli assegnamenti automatici, nella cartella *correttezza_assegnamenti*